## The Logic of Provability

David Hilbert[1] started a program he called "metamathematics," which treats mathematical proofs as objects of mathematical study. Previously, mathematicians studied such things as prime numbers, Fourier series, and conic sections. Now mathematicians study all those things and they also study mathematical proofs. The program was enormously frutiful. Its most dramatic results were the Gödel incompleteness theorems.[2] What we would like to do here is to conceptualize metamathematical results in terms of modal logic, interpreting "$\Box\ \varphi$" as saying that there is a proof of $\varphi$.

## Setting Things Up

The mathematical theories we'll talk about are all theories about the natural numbers, although the results we discuss are applicable much more widely. They apply to languages into which the basic laws of arithmetic are codable somehow. The *language of arithmetic* has as its nonlogical symbols "0", "S" (for successor), "+," "·," and "<," The main arithmetical theory we'll talk about is Peano Arithmetic (PA) whose axioms are:

$(\forall x) \sim Sx = 0.$

$(\forall x)(\forall y)(Sx = Sy \rightarrow x = y).$

$(\forall x)(x + 0) = x.$

$(\forall x)(\forall y)(x + Sy) = S(x + y)$.. 367-392.

$(\forall x)(x \cdot 0) = 0.$

---

[1] "On the Infinite" in Jean van Heijenoort, ed., *From Frege to Gödel* (Harvard University Press, 1967), pp

[2] "On Formally Undecidable Propositions in *Principia Mathematica* and Related Systems I" in *From Frege to Gödel*, pp. 592-617.

$(\forall x)(\forall y)(x \cdot Sy) = ((x \cdot y) + x)$.

$(\forall x) \sim x < 0$.

$(\forall x)(\forall y)(x < Sy \leftrightarrow (x < y \vee x = y))$.

Sentences (formulas without free variables) obtained form the following induction axioms schema by substituting an arithmetical formula for "R" and prefixing universal quantifiers to bind any free variables that result:

$((R0 \wedge (\forall x)(Rx \rightarrow RSx)) \rightarrow (\forall x)Rx)$

Virtually any argument that looks like ordinary, rigorous arithmetical reasonng can be formalized within PA. It's easy, once you get the knack.

We want to lay out some ideas from computability theory/ Where $\tau$ is an arithmetical term, we use $(\forall x < \tau)\varphi$ as an abbreviation for $(\forall x)(x < \tau \rightarrow \varphi)$ and $(\exists x < \tau)\varphi$ for $(\exists x)(x < \tau \wedge \varphi)$. "$(\forall x < \tau)$" and "$\exists x < \tau)$" are the bounded quantifiers. The bounded formulas are the smallest collection of expressions that contains the atomic formulas and is closed under "$\vee$," "$\wedge$," "$\sim$," and bounded quantification. The $\Sigma$ formulas are obtained by prefixing existential quantifiers to bounded formulas. A cornerstone of computability theory is the Church-Turing thesis: A partial function (a function from a subset of $\mathbb{N}^n$ to $\mathbb{N}$) is calculable iff it's the extension of a $\Sigma$ formula. A property or relation is effectively enumerable iff it's the extension of a $\Sigma$ formula.

Every true $\Sigma$ sentence is provable in PA; this is "$\Sigma$ completeness," which can be stated an proven within PA. Every effectively enumerable set S is weakly representable in PA, that is, there is a formula $\sigma(x)$ so that, for each n, PA $\vdash \sigma([n])$ iff $n \in S$. Here [n] is the numeral for n, the result of writing n "S"s before "0." Every decidable set R is stronly represntable in PA, that is, there is a formula $\rho(x)$ such that, for any n, in $n \in R$, PA $\vdash \rho([n])$, and if $n \notin R$, PA $\vdash \sim \rho([n])$.

Every calculable total function f is functionally representable in PA, that is, there is a formula $\mathscr{F}(x,y)$ such that, for any n, PA $\vdash (\forall y)(\mathscr{F}([n],y) \leftrightarrow y = [f(n)])$.

The breakthrough in metamathematics came with Gödel, who worked out the details of associating numerical codes with formulas and proof in such a way the important syntactic properties of expression of the language are correlated with decidable arithmetical properties of their codes. The number $\ulcorner \varphi \urcorner$ is the code of the formula φ. Where Γ is a recursively axiomatized theory (that is, there is an algorithm for telling whether a formula is an axiom), $\mathrm{Bew}_\Gamma([\ulcorner \varphi \urcorner])$ is the Σ sentence that says that there is a proof in Γ of φ. ("Beweis" is German for "proof.)

## The Incompleteness Theorems

The following result, which is the crucial move in Gödel's results, has always seemed a little bit magical to me:

> **Self-Reference Lemma.** For any formula ψ(x) there is a formula φ such that $(\varphi \leftrightarrow \psi([\ulcorner \varphi \urcorner])$ is a theorem of PA.

We can think of φ as talking about itself, saying "I have property ψ."

**Proof:** Let d be the following calculable function:

> If m has the form θ(x), $d(m) = \ulcorner \theta([m])$.
>
> Otherwise d(m) = 0.

d is calculable, so there is a formula $\mathscr{D}$ that functionally represents it. Let χ(x) be the formula $(\exists y)(\mathscr{D}(x,y) \wedge \psi(y))$, let $m = \ulcorner\chi(x)\urcorner$, and let $\varphi = \chi([m])$, so that $\ulcorner \varphi \urcorner = d(m)$. In PA, we can prove the following:

$$(\forall y)(\mathscr{D}([m],y) \leftrightarrow y = [\ulcorner \varphi \urcorner]).$$

$((\exists y)(\mathfrak{D}([m],y) \wedge \psi(y)) \leftrightarrow \psi([\ulcorner \varphi \urcorner]))$.

$(\varphi \leftrightarrow \psi([\ulcorner \varphi \urcorner]))$.⊠

Gödel actually wrote out the formula $\mathfrak{D}(x,y)$, rather than just claiming there has to be such a formula because d is calculable.

Use the Self-Reference Lemma to find a sentence γ, the "Gödel sentence," such that $(\gamma \leftrightarrow \sim Bew_{\Gamma}([\ulcorner \gamma \urcorner])$ is provable in PA. If there is a proof in Γ of γ, there is a proof in Γ of "$\sim Bew_{\Gamma}([\ulcorner \gamma \urcorner])$." But also, since "$Bew_{\Gamma}$" weakly represents provability in Γ, if there is a proof in Γ of γ, there is a proof in PA, and hence a proof in Γ, of "$Bew_{\Gamma}([\ulcorner \gamma \urcorner])$." So if γ is provable in Γ, Γ is inconsistent.

If the negation of γ is provable in Γ, then you can prove in Γ that there is a proof in Γof γ. But, unless Γ is inconsistent, there really isn't a proof in Γ, and for each n, you can prove that n isn't the code number of a proof in Γ of γ. A theory is said to be *ω-inconsistent* iff there is a formula θ(x) such that the theory proves θ([n]) for each n, but it also proves $(\exists x)\sim \theta(x)$. If γ is refutable in Γ, Γ is ω-inconsistent. This was the original version of Gödel's first incompleteness theorem: If Γ is ω-consistent, then there is a sentence that's undecidable in Γ, namely γ. Shortly afterward, Barclay Rosser[3] showed how to strengthen Gödel's result to this: In Γ is consistent, then there is a sentence that's undecidable in Γ.

We have a proof that is Γ is consistent, then γ isn't provable in Γ. We can restate this arithmetically, writing "CON(Γ)" to abbreviate "$\sim Bew_{\Gamma}([\ulcorner \sim 0 = 0 \urcorner])$":

$CON(\Gamma) \rightarrow \sim Bew_{\Gamma}(\ulcorner \gamma \urcorner)$.

[3]"Extensions of some Theorems of Gödel and Church," *Journal of Symbolic Logic* I (1936), pp. 231-235.

Hence,

$$\mathrm{CON}(\Gamma) \rightarrow \gamma.$$

We can formalize this as a proof in Γ. So if we had a proof in Γ of CON(PA), we could get a proof in Γ of γ, and we know this is only possible if Γ is inconsistent. This is Gödel's secondhenkin incompleteness theorem: No consistent, recursively axiomatized extension of PA can prove its own consistency.

## Löb's Theorem

Gödel showed that the sentence that says "I am not provable in PA" is true but not provable in PA. Leon Henkin[4] asked, "What about the sentence that says "I am provable in PA." M. H. Löb[5] answered his question: It is true and provable.

**Theorem (Löb)**. For Γ a recursively axiomatized extension of PA, $(\mathrm{Bew}([\ulcorner \varphi \urcorner]) \rightarrow \varphi)$ is provable in Γ if and only if φ is provable in Γ.

**Proof:** The right-to-left direction is obvious. Our method of proving the left-to-right is due to Kripke. Suppose φ isn't provable in Γ. Then $\Gamma \cup \{\sim \varphi\}$ is consistent. Then by the second incompleteness theorem, $\mathrm{CON}(\Gamma \cup \{\sim \varphi\})$ isn't provable in $\Gamma \cup \{\sim \varphi\}$. We can reformulate this arithmetically:

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

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

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

---

[4] "A Problem Concerning Provability," *Journal of Symbolic Logic* 17 (1952), p. 160.

[5] "Solution of a Problem of Leon Henkin," *Journal of Symbolic Logic* 20 (1955): 115-118.

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

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

**Translating Gödel and Löb's Results into Modal Terms**

Gödel and Löb discovered the main structural features of the logic of provability. We would like to explicate their discoveries in modal terms, interpreting "□" as "$\mathrm{Bew}_\Gamma$." Given Γ a recursively axiomatized extension of PA let's define a *Γ-interpretation* of our language for the modal sentential calculus is a function *i* that associates an arithmetical sentence with each modal formula, subject to the following constraints:

$i\,(\varphi \vee \psi) = (i(\varphi) \vee i(\psi))$

$i\,(\varphi \wedge \psi) = (i(\varphi) \wedge i(\psi))$

$i\,(\varphi \rightarrow \psi) = (i(\varphi) \rightarrow i(\psi))$

$i(\sim \varphi) = \sim i(\varphi)$

$i(\Box\, \varphi) = \mathrm{Bew}_\Gamma([\ulcorner i(\varphi) \urcorner])$

A modal formula φ is *always provable* in Γ iff, for each Γ-interpretation i, i(φ) is provable in Γ. φ is *always true* for Γ iff, for each Γ-interpretation i, i(φ) is true in the standard model $\mathbb{N}$.

Löb isolated three conditions that capture the structural features of provability that underlie Gödel's second incompleteness theorem. Löb's condition (L1) tells us that the set of always-provable formulas is closed under Necessitation:

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

(L2) tells us that the instances of (L1) are provable:

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

(L3) tells us that the instances of schema (K) are always provable:

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

Since the set of always-provable sentences is closed under (TC), we conclude that the set of always-provable sentences is a normal modal system that includes K4.

Löb's Theorem tells us that all instances of the following schema are always true:

(L) $\quad (\Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi)$

The proof of Löb's Theorem can be formalized in Γ, with the consequence that the instances of schema (L) are always provable. If we let GL (for "Gödel-Löb") be the smallest normal modal system that includes both (4) and (L), we see that the set of always-provable sentences includes GL. Dick de Jongh has shown that including schema (4) is redundant, so that GL can alternatively be characterized as the smallest normal modal system that includes (L).

The set of modal formulas always-provable in Γ is a normal modal system that includes GL. Is it, in fact, identical to GL? That depends on Γ. The second Gödel incompleteness theorem tells us that no consistent, recursively axiomatized theory proves its own consistency. So PA ∪ {~ CON(PA)} is consistent. If we take Γ to be PA ∪ {~ CON(PA)}, Γ, though consistent, contain a "proof" of its own inconsistency. "$\Box(P \wedge \neg P)$" is always provable in Γ, but it isn't in GL.

To get an exact matchup between being a theorem of GL and being always provable in Γ, we need to require more of Γ than just that it be consistent. Earlier we stipulated that an arithmetical theory is ω-inconsistent iff there is a formula φ(x) such that the theory proves φ([n]) but it also proves (∃x)~ φ(x). Let's say that the theory is *1-inconsistent* if the formula φ(x) is Σ.

Any inconsistent theory is 1-inconsistent, but PA ∪ {~ CON(PA)} is consistent but 1-inconsistent.

With this definition in hand, we can state the main theorem in provability logic, which is due to Robert Solovay[6]:

> **Theorem** (Solovay). If Γ is a 1-consistent, recursively axiomatized extension of PA, a modal formula is always provable in Γ if and only if it's in GL.

The right-to-left direction follows from (L1), (L2), (L3), and Löb's theorem. Proving the left-to-right direction will be a major undertaking. Before setting out, let's spend a little time exploring the relation between "□" and "Bew." Take Γ to be a recursively axiomatized extension of PA. Until further notice, we'll take "Bew" to mean "$\mathrm{Bew}_\Gamma$"; "always provable" will mean "always provable in Γ." Define the *arithmetical frame* for Γ to be the pair <W,R> where W is the collection[7] of models of Γ and we stipulate that wRv iff, whenever Bew([⌜ψ⌝]) is true in w, ψ is true in v. We have the following:

> **Theorem**. For any arithmetical sentence φ and world w, Bew([⌜φ⌝]) is true in w if and only if φ is true in every world accessible from w.

**Proof: (⇒)** is immediate. To get **(⇐)**, suppose that Bew([⌜φ⌝]) isn't true in w. We want to construct a world accessible from w in which φ is false. In other words, we want to construct a model of Γ ∪ {¬ φ} ∪ {ψ: Bew([⌜ψ⌝]) is true in w}. Suppose, for *reductio*, that there is no such

[6] "Provability Interpretations of Modal Logic," *Israel Journal of Mathematics* 25 (1976): 287-304. The definitive exposition of provability logic is George Boolos, *The Logic of Provability* (Cambridge: Cambridge University Press, 1995).

[7] A technical issue we'll ignore is that, in term of standard set theory, W won't be a set, but a proper class.

model. Then by compactness, there exist sentences $\psi_1, \psi_2, \ldots \psi_n$, with each Bew([⌜$\psi_i$⌝]) true in w, such that $\Gamma \cup \{\neg \varphi\} \cup \{\psi_1, \psi_2, \ldots, \psi_n\}$ is in inconsistent. Consequently,

$$\Gamma \vdash (\psi_1 \rightarrow (\psi_2 \rightarrow \ldots \rightarrow (\psi_n \rightarrow \varphi)\ldots)).$$

By (L1),

$$\Gamma \vdash \mathrm{Bew}([\ulcorner(\psi_1 \rightarrow (\psi_2 \rightarrow \ldots \rightarrow (\psi_n \rightarrow \varphi)\ldots))\urcorner]).$$

Using (L3) multiple times:

$$\Gamma \vdash (\mathrm{Bew}([\ulcorner\psi_1]) \rightarrow (\mathrm{Bew}([\ulcorner\psi_2\urcorner]) \rightarrow \ldots \rightarrow (\mathrm{Bew}(\ulcorner\psi_n\urcorner]) \rightarrow \mathrm{Bew}([\ulcorner\varphi\urcorner]))\ldots)).$$

This conditional and each of the Bew([⌜$\psi_i$⌝])s are true in w. So Bew([⌜$\varphi$⌝]) is true in w. Contradiction.⊠

The worlds are those models of the language or arithmetic in which all the theorems of $\Gamma$ are true. If Bew([⌜$\varphi$⌝]) is true in w, then w thinks that $\varphi$ is a theorem of $\Gamma$. If v is a model of the language of arithmetic in which $\varphi$ is false, w won't regard v as a genuine world. This can happen even if Bew([⌜$\varphi$⌝]), though true in w, is false in the standard model and v really is a possible world. A possible world – that is, model of $\Gamma$– is accessible from w only if w recognizes it as a model of $\Gamma$.

Given an arithmetical interpretation i of the modal language, define an interpretation $I_i$ of the arithmetical frame <W,R> by stipulating that, for $\alpha$ an atomic modal formula, $I_i(\alpha,w) = 1$ iff $i(\alpha)$ is true in w. The theorem enables us to prove that, for any arithmetical interpretation i and modal formula $\varphi$, $\varphi$ is true in the model $\langle W,R,I_i,w\rangle$ iff $i(\varphi)$ is true in the model w for the language of arithmetic. $\varphi$ is always provable iff for every arithmetical interpretation i and world w, $\varphi$ is true in $\langle W,R,I_i,w\rangle$, which happens iff, for every arithmetical interpretation i and world w, $i(\varphi)$ is true in w.

If Γ is true, then the standard model is one of the worlds in W, and it has access to itself and every other world. A modal formula φ is always true iff, for every arithmetical interpretation i, φ is true in <W,R,$I_i$,ℕ>, which happens iff, for every arithmetical interpretation i, i(φ) is true in ℕ.

## Possible-world Semantics for GL

We now want to set all the arithmetical stuff aside for a while and just think about GL as a system of axioms for the modal sentential calculus. The aim is to understand GL in terms of possible-world semantics. We begin with a definition and a theorem:

> **Definition.** A binary relation R on W is *well-capped* iff, any nonempty subset X of W has an R-maximal element, that is, an element of X that doesn't bear R to any element of x.

Equivalently, <W,R> is well-capped iff there is no infinite R-sequence $w_0$ R $w_1$ R $w_2$ R $w_3$ R....

> **Theorem.** Assuming "P" is an atomic sentence of our modal language, the following are equivalent, for any uninterpreted frame <W,R>:

(i) R is transitive and well-capped.

(ii) All the sentences in GL are valid for <W,R>.

(iii) The formulas "(□P → □□P)" and "(◊P → ◊(P ∧ ¬◊P))" are valid for <W,R>.

**Proof: (i) implies (ii):** We already know that, if R is transitive, all the instances of (4) are valid for <W,R>. We now want to see that, if R is transitive and well-capped, all the instance of (L) are valid for <W,R>. You can prove in K that (L) is equivalent to (◊¬φ → ◊(¬φ ∧ ¬◊¬φ)).

Suppose $\Diamond \neg \varphi$ is true in world w. Let X be the set for worlds accessible from w in which $\neg \varphi$ is true. The X has an R-maximal element v. v is a world accessible from w in which $(\neg \varphi \wedge \neg \Diamond \neg \varphi)$ is true. Thus we know that the set of sentence valid for <W,R> is a normal modal system that includes (4) and (L). It follows that it includes GL.

**(ii) implies (iii):** Immediate, since the formulas are both in GL.

**$\neg$ (i) implies $\neg$ (iii).** We have already seen how, if R isn't transitive, we can find a model <W,R,I,@> in which "$(\Box P \rightarrow \Box \Box P)$" is false. Let's suppose that R isn't well-capped, so that there is an infinite R-sequence $w_0$ R $w_1$ R $w_2$ R $w_3$ R.... Define the interpretation I so that I("P",x) = 1 iff x is one of the $w_i$s; it doesn't matter what I does with the other atomic sentences. Because "P" is true in $w_1$, "$\Diamond$P" is true in $w_0$. Any world accessible from $w_0$ in which "P" holds will be $w_i$ for some i, and it will have access to a world in which "P" holds, namely $w_{i+1}$. So "$\Diamond(P \wedge \neg \Diamond P)$" is false in w.⊠

We now want to show that a sentence is in GL if and only if it's true in every model in which the accessibility relation is transitive and well-capped. We just proved the left-to-right direction (soundness). We still have to prove the right-to-left direction (completeness).

We have a standard method for proving completeness theorems for normal modal systems, namely, looking at the canonical frame. This method won't work with GL. GL being normal, it has a canonical frame <W,R,I>. If $\chi$ isn't in GL, then there is a world @ in the canonical frame in which $\chi$ is false. So far, so good, but now the proof breaks down. The canonical frame of <G,L> isn't well-capped. To see this, pick a PA-interpretation i, and consider the arithmetical frame <W,R> for PA. ℕ is a world in W that has access to every world. In particular, it has access to itself. So the set of formulas true in <W,R,$I_i$,ℕ> includes all the formulas of the form ($\Box \varphi \rightarrow \varphi$) and also all the formulas in GL. So GL $\cup$ {instances of (T)} is truth-functionally consistent. So there is a world @ in the canonical frame for GL that includes all the instances of (T). We have @ R @ R @ R @ R ....

We can still prove the completeness theorem, but we have to do it in a roundabout way, by showing the following are equivalent:

(i) $\chi$ is in GL.

(ii) $\chi$ is true in every model in which the accessibility relation is transitive and well-capped.

(iii) $\chi$ is true in every model in which the accessibility relation is a finite partial order, that is, a finite relation that is antireflexive (no world has access to itself) and transitive.

Actually, it will be useful to prove something further. Define a *tree* to be a partial order that meets two extra conditions:

There is a least member.

Branch-connection: If Ruw and Rvw, then either Ruv or u=v or Rvw.

The paradigm case of a finite true is a nonempty, finite set of finite sequences, with the property that every initial segment of a member of the set is a member of the set. Ruv holds if and only if v extends u. A *tree model* is a model <W,R,I,@> in which <W,R> is a tree with @ as its least element.

**Theorem.** For $\chi$ a modal formula, the followin are equivalent:

(i) $\chi$ is in GL.

(ii) $\chi$ is true in every model in which the accessibility relation is transitive and well-capped.

(iii) $\chi$ is true in every model in which the accessibility relation is a finite partial order.

(iv) $\chi$ is true in every finite tree model.

**Proof:** We've already shown that (i) implies (ii). It's evident that (ii) entails (iii) and (iii) entails (iv). We want to show that the negation of (i) entails the negation of (iii). Later we'll worry about the path from (iv) to (iii).

Suppose that $\chi$ isn't in GL. We want to construct a finite tree model <W,R,I,@> in which $\chi$ is false. The construction we've used in the past, with maximal consistent sets of sentences, won't give us a finite tree. To keep everything finite, we don't look at all the sentences, but only at the sentences that are either subsentences of $\chi$ or negations of subsentences of $\chi$. Since $\chi$ isn't an element of GL, we can find a set of sentences @ with the following properties:

$\neg\chi$ is an element of @.

@ is GL-consistent.

Every member of @ is either a subsentence of $\chi$ or the negation of a subsentence of $\chi$.

For each subsentence of $\chi$, either it or its negation is in @.

To form @, we go through the subsentences of $\chi$. When we come to a sentence, we add either it or its negation to our set, preserving GL-consistency.

Let W be the set of all maximal GL-consistent sets of subsentences of $\chi$ and negated subsentences of $\chi$. That is, a set of sentences is in W iff it meets the last three of the four conditions above. If w is an element of W and $\alpha$ is an atomic sentence that occurs in $\chi$, we'll set

$I(\alpha,w) = 1$ iff $\varphi \in w$. The tricky part is defining the accessibility relation R. Here's the definition: w R v iff the following two conditions are met:

For any sentence $\Box\varphi$ that's an element of w, both $\Box\varphi$ and $\varphi$ are elements of v.

There is a sentence $\theta$ such that $\Box\theta$ is in v, but $\Box\theta$ isn't in w.

This stipulation makes $\langle W,R\rangle$ into a finite partial order. The second clause is a brute force method of guaranteeing antireflexivity.

The proof that, for any sentence $\psi$ that's a subsentence of $\chi$, $\psi$ is true in w in the model $\langle W,R,I,@\rangle$ iff $\psi$ is an element of w is routine, except for one part. We need to show that, if $\Box\psi$ is a subsentence of $\chi$ that isn't in w, then there is a v with wRv that doesn't contain $\psi$. So v will have to contain $\sim\psi$ and all the sentences $\varphi$ and $\Box\varphi$ with $\Box\varphi$ in w, and it will also have to contain some sentence that begins with a "$\Box$" that's isn't in w. The sentence we'll use for this purpose is $\Box\psi$.To prove the existence of such a v, we need to show that $\{\sim\psi, \Box\psi\} \cup \{(\varphi \wedge \Box\varphi): \Box\varphi \in w\}$ is GL-consistent. Suppose it isn't, so that, for some $\varphi_1, \varphi_2, \ldots, \varphi_n$ with each $\Box\varphi_i$ in w, GL implies:

$$((\Box\varphi_1 \wedge \varphi_1) \rightarrow ((\Box\varphi_2 \wedge \varphi_2) \rightarrow \ldots \rightarrow ((\Box\varphi_n \wedge \varphi_n) \rightarrow (\Box\psi \rightarrow \psi))\ldots)),$$

Because GL is normal, the following sentence is in GL:

$$(\Box(\Box\varphi_1 \wedge \varphi_1) \rightarrow (\Box(\Box\varphi_2 \wedge \varphi_2) \rightarrow \ldots \rightarrow (\Box(\Box\varphi_n \wedge \varphi_n) \rightarrow \Box(\Box\psi \rightarrow \psi))\ldots)).$$

Because GL includes K4, $(\Box\varphi_i \rightarrow \Box(\Box\varphi_i \wedge \varphi_i))$ is in GL, for each i, and also, because GL includes (L), $(\Box(\Box\psi \rightarrow \psi) \rightarrow \Box\psi)$ is in GL. Consequently, the following sentence is in GL:

$$(\Box\varphi_1 \rightarrow (\Box\varphi_2 \rightarrow \ldots \rightarrow (\Box\varphi_n \rightarrow \Box\psi)\ldots)).$$

Because each of the $\Box\varphi_i$s is in w, $\Box\psi$ is in w. Contradiction.

We still need to show that (iii) implies (iv). Suppose that R is a finite partial order on a set W and that $\chi$ is false in <W,R,I,@>. Let $W^\dagger$ be the set of all finite sequences beginning with @ with the property that each member of the sequence bears R to the next member of the sequence, if there is one. Sequence s bears $R^\dagger$ to sequence t iff s is a proper initial segment of t. $I^\dagger(\alpha,s) = I(\alpha,$ the last member of s). An easy induction shows that, for each formula $\varphi$ and sequence s in $W^\dagger$, $\varphi$ is true in $<W^\dagger,R^\dagger,I^\dagger,s>$ iff $\varphi$ is true in <W,R,I, last member of s>. In particular, $\chi$ is false in the finite tree model $<W^\dagger,R^\dagger,I^\dagger,<@>>$.⊠

This theorem gives us a decision procedure for GL. If a sentence is in GL, we can derive it, whereas if a sentence is outside GL the proof of the theorem shows us how to construct a finite tree model in which it's false.

**GL = {Always-provable Formulas}**

Now we're ready for the big time. Given a sentence $\chi$ that's not in GL, we want to find a $\Gamma$-nterpretation i such that i($\chi$) isn't a consequence of $\Gamma$. We can find a finite tree model <W,R,I,@> in which $\chi$ is false. We can emumerate the members of W as $w_1, w_2,..., w_n$, listed in such a way that $i < j$ whenever $w_i \, R \, w_j$. Thus $@ = w_1$. We expand the model by adding $w_0$ as an extra world, stipulating that every other world is accessible from $w_0$ and that $I(\alpha,w_0) = I(\alpha,w_1)$, for $\alpha$ atomic. At the end of the day, when we get our $\Gamma$-interpretation, world $w_0$ will play the role of the actual world, that is, the standard model. The sentences true in world $w_1$ might or might not be true in the standard model; we don't want to presume. When we turn to the logic of almost-truth, world $w_0$ will play a starring role.

Our plan is looking for a Γ-interpretation that reproduces the structure of the tree is reminiscent of the strategy we use in plain sentential calculus to see how to find an SC sentence with a given truth table. What we did there was to find, for each line of a truth table, a sentence, the state description, that described that line, then to take our sentence to be the disjunction of the state descriptions of the lines at which the given truth table assigns the value "true." Pursuing the same plan here, we want to find, for each world $w_j$, a sentence $\sigma_j$ that describes that world. Once we've done that, we can take our arithmetical interpretation to be the function that assigns to each atomic formula the disjunction of the world-descriptions of the worlds in which the formula is true. Specifically, we find, for each $j \leq n$, a sentence $\sigma_j$ meeting these conditions:

(i) $\mathrm{PA} \vdash (\bigvee_{0 \leq j \leq n} \sigma_j)$. ("$\bigvee$" denotes a disjunction.)

(ii) $\mathrm{PA} \vdash \sim (\sigma_j \wedge \sigma_k)$, for $j \neq k$.

(iii) $\mathrm{PA} \vdash (\sigma_j \rightarrow \sim \mathrm{Bew}([\ulcorner \sim \sigma_k \urcorner]))$, whenever $w_j \mathrm{R}\, w_k$.

(iv) $\mathrm{PA} \vdash (\sigma_j \rightarrow \mathrm{Bew}([\ulcorner \bigvee_{w_j R w_k} \sigma_k \urcorner]))$, for $1 \leq j \leq n$.[8]

(v) $\mathrm{PA} \vdash (1\text{-}\mathrm{CON}(\Gamma) \rightarrow \sigma_0)$.

Defining our Γ-interpretation i by stipulating that, for φ atomic, i(φ) is the disjunction of the $\sigma_j$s, for j a world in which φ is true, gives us the following:

**Claim.** For any j, $1 \leq j \leq n$, and any modal formula φ, if φ is true in $w_j$, then $\mathrm{PA} \vdash (\sigma_j \rightarrow i(\varphi))$.

**Proof:** We prove by induction on the complexity of formulas that, for each formula φ, if φ is true in $w_j$, then $\mathrm{PA} \vdash (\sigma_j \rightarrow i(\varphi))$, whereas if φ is false in $w_j$, $\mathrm{PA} \vdash (\sigma_j \rightarrow \neg\, i(\varphi))$. If φ is atomic, then if φ

[8] In case there aren't any worlds accessible from j, let me stipulate that I'll take the "disjunction" of the $\sigma_j$s with Rjk to be the logically inconsistent sentence "$\sim 0 = 0$." So (iv) tells us that, if there aren't any world accessible form j, $\mathrm{PA} \vdash (\sigma_j \rightarrow \sim \mathrm{Con}(\Gamma))$.

is true in $w_j$, $\sigma_j$ is one of the disjuncts of i(φ), whereas, if φ is false in $w_j$, condition (ii) assures us that $\sigma_j$ is provably incompatible with each of the disjuncts of $\sigma_j$. In case φ is built up from simpler formulas by means of the SC connectives, the proof is easy and I won't go through it here. Here let's worry instead about showing that the claim holds when φ has the form □ψ.

Let's say the worlds accessible from $w_j$ are $w_{k_1}$, $w_{k_2}$,..., $w_{k_m}$. If □ψ is true in $w_j$, then by inductive hypothesis, for each h, $1 \leq h \leq m$, PA $\vdash (\sigma_{k_h} \rightarrow i(\psi))$. So PA $\vdash ((\sigma_{k_1} \vee \sigma_{k_2} \vee ... \vee \sigma_{k_m}) \rightarrow i(\psi))$. By (L1) and (L3), PA $\vdash$ (Bew([⌜$(\sigma_{k_1} \vee \sigma_{k_2} \vee ... \vee \sigma_{k_m})$⌝]) → Bew([i(ψ)]). Since, by (iv),[9] PA $\vdash$ ($\sigma_j$→ Bew([⌜$(\sigma_{k_1} \vee \sigma_{k_2} \vee ... \vee \sigma_{k_m})$⌝])), PA $\vdash$ ($\sigma_j$ → i(□ψ)).

If, on the other hand, □ψ is false in $w_j$, then there is a world $w_k$ accessible from $w_j$ in which ψ is false. By inductive hypothesis, PA $\vdash$ ($\sigma_k$ → ¬ i(ψ)). It follows by (L1) and (L3) that Γ $\vdash$ (Bew([⌜i(ψ)⌝]) → Bew([⌜¬$\sigma_k$⌝])), and so PA $\vdash$ (¬ Bew([⌜¬$\sigma_k$⌝]) → ¬ i(□ψ)). It follows by (iii) that PA $\vdash$ ($\sigma_j$ →¬ i(□ψ)).⊠

Given the Claim, we know that PA $\vdash$ ($\sigma_1$ → ¬ i(χ)). It follows by (L1) and (L3) that PA $\vdash$ (Bew([⌜i(χ)⌝]) → Bew([⌜¬ $\sigma_1$⌝])), and so, by (iii), PA $\vdash$ ($\sigma_0$ → ¬ Bew([⌜i(χ)⌝])). Since, by (v) and the hypothesis that Γ is 1-consistent, $\sigma_0$ is true, it follows that Bew([⌜i(χ)⌝]) is false, so that i(χ) isn't a consequence of Γ.

It remains to find the $\sigma_j$s. Figuring out what formulas to write down took a lot of ingenuity of Solovay's part, and I won't attempt to motivate the construction. I'll just write the formulas down and verify that they work. Define a function f(x,y) as follows:

If z isn't the Gödel number of a formula whose only free variable is "x," f(y,z) = 0.

[9]This is where the proof gets stuck for j = 0, since (iv) only applies where $1 \leq j \leq n$. When we turn to the logic of always true formulas, we'll develop a restricted version of the Claim that applies to world $w_0$.

Suppose that z is the Gödel number of a formula ψ(x) with "x" as its only free variable We define f(y,z) by induction on x:

f(0,z) = 0.

If f(m,z) = j and m is a proof in Γ that ends in ψ([k]) with $w_j$ R $w_k$, then f(m+1,z) = k.

Otherwise, f(m+1,z) = f(m,z).

If z = ⌜ψ(x)⌝, then in calculating the value of f(y,z) for different values of y, we start at node $w_0$ and make our way up the tree. If, at a certain point, we're at node $w_j$ and we find a proof of ⌜ψ[k])⌝, with $w_j$ R $w_k$, then we jump to node $w_k$. Because the tree is finite, the jumping will have to come to a halt eventually.

f is a calculable total function. So we can find encode the recursive definition of f as a Σ explicit definition, and having done so, we can prove the basic features of f in PA. We can prove that, for each number p, λyf(y,p), the function taking y to f(y,p), is a nondecreasing total function whose range is a subset of {0,1,2,..., n}. The plan is to find a particular number p and take $\sigma_j$ to be the sentence, "j is the greatest value output by λy f(y,p)." If p isn't the code of a formula with "x" as its only free varible, λy f(y,p) will be the constant function 0, so identifying $\sigma_j$ with "j is the greatest value output by λy f(y,p)" will give us (i), (ii), (iv), (v), but not (iii). To get all five conditions, we'll need the second argument of f to be the code of a formula with "x" as its only free variable. We want to find an appropriate ψ(x) so that taking $\sigma_j$ to be "j is the greatest value output by λy f(y, ⌜ ψ(x) ⌝)" gives us the five conditions.

Take any formula ψ(x) with "x" as its only free variables.

(i*) PA $\vdash \bigvee_{0\leq j\leq n}$ [j] is the greatest value output by λy f(y,[⌜ ψ(x) ⌝]).

(ii*) For $j \neq k$, PA $\vdash \neg$ ([j] is the greatest value output by $\lambda y\ f(y,[\ulcorner \psi(x) \urcorner]) \wedge$ [k] is the greatest value output by $\lambda y\ f(y,[\ulcorner \psi(x) \urcorner])$).

If the greatest value output by $\lambda y\ f(y, \ulcorner \psi \urcorner)$ is j, then there is a least number $y_0$ with $f(y_0, \ulcorner \psi(x) \urcorner) = j$. There is no number $y_1$ greater that $y_0$ that codes a proof whose last line is $\psi([k])$ with $w_j$ R $w_k$, since if there were such a number, $f(y_1, \ulcorner \psi(x) \urcorner)$ would be equal to k. If there were a proof of $\psi([k])$, then there would be a number greater than $y_0$ that codes a proof that ends in $\psi([k])$. The reason will be familiar to anyone who has stretched out a seven-page paper to fulfill a requirement of a ten-page term paper. You pad the paper by adding extra stuff. If you have a proof of $\psi([k])$, you can write $\ulcorner \psi([k]) \urcorner$ over and over at the end of the argument, getting proofs with larger and larger Gödel numbers until you get one whose Gödel number is greater than $y_0$. We can formalize this reasoning in PA, getting this:

(iii*) PA $\vdash$ ([j] is the greatest value output by $\lambda y\ f(y,[\ulcorner \psi(x) \urcorner]) \rightarrow \sim \text{Bew}([\ulcorner \psi([k])\urcorner]))$ , for $w_j$ R $w_k$.

We also have this:

PA $\vdash$ ( [j] is the greatest value output by $\lambda y\ f(y,[\ulcorner\psi(x) \urcorner]) \rightarrow (\exists y)f(y,[\ulcorner \psi(x) \urcorner]) = [j]$).

$\Sigma$-completeness gives us this:

PA $\vdash ((\exists y)f(y,[\ulcorner \psi(x) \urcorner]) = [j] \rightarrow \text{Bew}([\ulcorner (\exists y)f(y,[\ulcorner \psi(x) \urcorner]) = [j] \urcorner]))$.

Once the function gets to j it either stays there or moves on to some value k with $w_j$ R $w_k$:

PA $\vdash ((\exists y)f(y, [ \ulcorner \psi(x) \urcorner ]) = [j] \rightarrow$ ([j] is the greatest value output by $\lambda y\ f(y, [\ulcorner \psi(x) \urcorner] \vee (\bigvee_{w_j R w_k}$ [k] is the greatest value output by $\lambda y\ f(y,[\ulcorner \psi(x) \urcorner]))))$.

(L1) and (L3) let us place a "Bew" in front of this and put it through:

$$PA \vdash (Bew([\ulcorner(\exists y)f(y, [\ulcorner \psi(x) \urcorner] = [j] \urcorner]) \rightarrow Bew([\ulcorner([j] \text{ is the greatest value output by } \lambda y\ f(y, [\ulcorner \psi(x) \urcorner] \vee (\bigvee_{w_jRw_k}[k] \text{ is the greatest value output by } \lambda y\ f(y,[\ulcorner \psi(x) \urcorner])) \urcorner]))).$$

Putting these results together, we get:

(iv*) $$PA \vdash ([j] \text{ is the greatest value output by } \lambda y\ f(y,[\ulcorner \psi \urcorner]) \rightarrow Bew([\ulcorner([j] \text{ is the greatest value output by } \lambda y\ f(y, [\ulcorner \psi(x) \urcorner] \vee (\bigvee_{w_jRw_k}[k] \text{ is the greatest value output by } \lambda y\ f(y,[\ulcorner \psi(x) \urcorner])) \urcorner]))).$$

Now we're ready to put the pieces together. Use the Self-Reference Lemma to find a formula $\sigma(x)$ such that

$$PA \vdash (\forall z)(\sigma(z) \leftrightarrow z \text{ is the greatest value output by } \lambda y\ f(y,[\ulcorner \neg \sigma(x) \urcorner])).$$

Rewriting $\sigma([j])$ as $\sigma_j$, (i*), (ii*), and (iii*) give us (i), (ii), and (iii). Let's think about (iv). (iv*) gives us this:

$$PA \vdash (\sigma_j \rightarrow Bew([\ulcorner (\sigma_j \vee (\bigvee_{w_jRw_k} \sigma_k)) \urcorner])).$$

$\sigma_j$ implies (in PA) that j is the greatest value output by $\lambda y\ f(y, \ulcorner \neg \sigma(x) \urcorner)$, which implies that, for some y, $f(y,\ulcorner \neg \sigma(x) \urcorner) = j$, which implies (given that $j \neq 0$) that y is the code of a proof of $\neg \sigma_j$. Formalizing this, we get:

$$PA \vdash (\sigma_j \rightarrow Bew([\ulcorner \neg \sigma_j \urcorner])).$$

We get (iv) by formalizing an argument of the form:

$(P \vee Q)$

$\neg P$

$\therefore Q$

Finally, we want to prove (v),[10] that is, we want to show that, provided $\Gamma$ is 1-consistent, the greatest element of $\{f(y,\ulcorner\sim\sigma(x)\urcorner): y \in \mathbb{N}\}$ is 0. Take j with $1 \leq j \leq n$. (iv) tells us:

$$PA \vdash (\sigma_j \rightarrow Bew([\ulcorner \bigvee_{w_jRw_k} \sigma_k \urcorner])).$$

For each k, $1 \leq k \leq n$, we have

$$PA \vdash (\sigma([k]) \rightarrow (\exists y)f(y,[\ulcorner\sim\sigma(x)\urcorner] = [k])$$

Therefore,

$$PA \vdash ((\bigvee_{w_jRw_k} \sigma_k) \rightarrow (\bigvee_{w_jRw_k} (\exists y)f(y,[\ulcorner \neg \sigma(x)\urcorner]) = [k])).$$

$$PA \vdash ((\bigvee_{w_jRw_k} \sigma_k) \rightarrow (\exists y)( \bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x)\urcorner]) = [k])).$$

Applying (L1) and (L3) yields:

$$PA \vdash (Bew([\ulcorner \bigvee_{w_jRw_k} \sigma_k \urcorner]) \rightarrow Bew([\ulcorner (\exists y)( \bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x)\urcorner]) = [k] \urcorner]))).$$

Using (iv),

($) $$PA \vdash (\sigma_j \rightarrow Bew([\ulcorner (\exists y)( \bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x)\urcorner]) = [k] \urcorner]))).$$

$\sigma_j$ tells us that j is the greatest value output by $\lambda y\ f(y, \ulcorner \neg \sigma(x) \urcorner)$, so that, if $w_j$ R $w_k$, there isn't any y with $f(y, \ulcorner \neg \sigma(x) \urcorner) = k$. Formalizing this reasoning in PA, we get, for each k with $w_j$ R $w_k$,

$$PA \vdash (\sigma_j \rightarrow \neg (\exists y)f(y,[\ulcorner \neg \sigma(x) \urcorner]) = [k]).$$

$$PA \vdash (\sigma_j \rightarrow \neg (\exists y)(\bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x) \urcorner]) = [k]).$$

$$PA \vdash (\sigma_j \rightarrow (\forall y) \neg (\bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x) \urcorner]) = [k]).$$

$\Sigma$-reflection tells us the following, since $\neg (\bigvee_{w_jRw_k} f(y,[\ulcorner \neg \sigma(x) \urcorner]) = [k])$ is provably equivalent to a $\Sigma$ formula:

[10] If we were assuming that $\Gamma$ were true, instead of merely that it's $\omega$-consistent, proving (v) would be a piece of cake. For j > 0, $\sigma([j])$ asserts its own refutability, so that, if it were true, it would be a true refutable sentences. However, we are not assuming that $\Gamma$ is true, so there may well be true sentences that are refutable in $\Gamma$. So we have more work to do.

PA $\vdash (\forall y)(\neg(\bigvee_{w_j R w_k} f(y,[\ulcorner \neg \sigma(x) \urcorner]) = [k]) \rightarrow Bew([\ulcorner \neg (\bigvee_{w_j R w_k} f([y],[\ulcorner \neg \sigma(x) \urcorner])$
$= [k]) \urcorner]))$.

Putting these together, we get:

(¢) PA $\vdash (\sigma_j \rightarrow (\forall y)\, Bew([\ulcorner \neg (\bigvee_{w_j R w_k} f([y],[\ulcorner \neg \sigma(x) \urcorner]) = [k]) \urcorner])))$.

($) and (¢) yield:

PA $\vdash (\sigma_j \rightarrow \neg$ 1-CON$(\Gamma))$.

This holds for every j, $1 \leq j \leq n$. Hence by (i):

PA $\vdash (\neg \sigma_0 \rightarrow \neg$ 1-CON$(\Gamma))$.

This is the contrapositive of (v).⊠

**GLS = {Always-true Formulas}**

We now turn out attention to problem of determining which modal formulas are always true. Assuming that Γ is true, every always-provable formula will be always true, but not every always-true formula will be always provable, for all the instances of schema (T) will be always-true, but only those with always-provable consequents will be always provable. It turns out that these two observations, together with the recognition that the always-true formulas are closed under tautological consequence, is enough to give us a complete inventory of always-true formula.

Let GLS (for "Gödel-Löb-Solovay") be the smallest collection of formulas that includes GL and all the instances of schema (T) and is closed under tautological consequence.

**Theorem (Solovay).** Given a modal formula $\chi$, let the subformulas of $\chi$ that begin with "□" be $\Box\eta_1, \Box\eta_2,..., \Box\eta_m$. The following are equivalent:

① $\chi \in$ GLS.

② $(((\Box\eta_1 \rightarrow \eta_1) \wedge (\Box\eta_2 \rightarrow \eta_2) \wedge ... \wedge (\Box\eta_m \rightarrow \eta_m)) \rightarrow \chi) \in$ GL.

③ $\chi$ is always true.

**Proof:** That ② implies ① and that ① implies ③ are obvious (given Löb's theorem), so all we need to show is that ③ implies ②. Actually, we'll show that the negation of ② implies the negation of ③. If the conditional $(((\Box\eta_1 \rightarrow \eta_1) \wedge (\Box\eta_2 \rightarrow \eta_2) \wedge ... \wedge (\Box\eta_m \rightarrow \eta_m)) \rightarrow \chi)$ isn't in GL, we follow the same procedure as before to find a model $<\{w_0, w_1,..., w_n\},R,I,w_0>$ in which the conditional is false at world $w_1$. We want to show that a subformula of $\chi$ is true in world $w_0$ if and only if it's true at world $w_1$. For atomic formulas, this follows immediately from the way we, thinking ahead, stipulated truth values when extending the model to include world $w_0$. For conjunctions, disjunctions, conditionals, biconditionals, and negations, the proof is easy. If $\Box\eta_j$ is true in world $w_0$, then $\eta_j$ is true in every world accessible from $w_0$. Since every world accessible from world $w_1$ is accessible from world $w_1$, it follows that $\eta_j$ is true in every world accessible from world $w_1$, and so $\Box\eta_j$ is true in world $w_1$. If, on the other hand, $\Box\eta_j$ is true in world $w_1$, then $\eta_j$ is true in every world accessible from world $w_1$. The only world accessible from $w_0$ that isn't accessible from $w_1$ is $w_1$ itself. Since $(\Box\eta_j \rightarrow \eta_j)$ is true in $w_1$, $\eta_j$ is true in true in $w_1$, and thus true in every world accessible from $w_0$, so that $\Box\eta j$ is true in $w_0$.

In particular, since $\chi$ is false in $w_1$, $\chi$ is false in $w_0$.

We now want to show that, for each subsentence $\theta$ of $\chi$, if $\theta$ is true in $w_0$, PA $\vdash (\sigma_0 \rightarrow i(\theta))$, whereas if $\theta$ is false in $w_0$, PA $\vdash (\sigma_0 \rightarrow \neg\, i(\theta))$. Since $\sigma_0$ is true, it will follow that $i(\chi)$ is false, as required.

The proof for $\theta$ atomic is the same as the proof we gave earlier for worlds 1, 2,..., n. The proof for $\theta$ a disjunction, conjunction, conditional, biconditional, or negation is routine.

Suppose that $\Box\eta_j$ is true in $w_0$. For each $k > 0$, $w_k$ is accessible from $w_0$, and so $\eta_j$ is true in $w_k$. We showed earlier that this shows that PA $\vdash (\sigma_k \rightarrow i(\eta_j))$. Since $\eta_j$ is true in $w_1$ and the same subsentences of $\chi$ are true in $w_0$ and in $w_1$, $\eta_j$ is true in $w_0$, and so, by inductive hypothesis, PA $\vdash (\sigma_0 \rightarrow i(\eta_j))$. It follows that PA $\vdash ((\sigma_0 \vee \sigma_1 \vee ... \vee \sigma_n) \rightarrow i(\eta_j))$. Since PA $\vdash (\sigma_0 \vee \sigma_1 \vee ... \vee \sigma_n)$, we have PA $\vdash i(\eta_j)$. By (L1), PA$\vdash$ Bew([$\ulcorner i(\eta_j)\urcorner$]), that is, PA $\vdash i(\Box\eta_j)$, and so PA $\vdash (\sigma_0 \rightarrow i(\Box\eta_j))$.

Now suppose instead that $\Box\eta_j$ is false in $w_0$. Then there is a world $w_k$ with $k > 0$ in which $\eta_j$ is false. We showed earlier that this implies that PA $\vdash (\sigma_k \rightarrow \neg\, i(\eta_j))$. Applying contraposition, (L1), (L3), and contraposition again, we obtain PA $\vdash (\neg$Bew([$\ulcorner\neg\sigma_k\urcorner$]) $\rightarrow \neg i(\Box\eta_j))$. Since (iii) gives us PA $\vdash (\sigma_0 \rightarrow$ $\neg$Bew([$\ulcorner\neg\sigma_k\urcorner$])), it ollows that PA $\vdash (\sigma_0 \rightarrow \neg i(\Box\eta_j))$.⊠

From the equivalence of ② and ③ and the existence of a decision procedure for GL, we see that there is an algorithm of testing whether a modal formula is always true.