Logics of (formal and informal) provability

Rafal Urbaniak
Centre for Logic and Philosophy of Science, Ghent University
Institute of Philosophy, Sociology and Journalism, University of Gdansk
rfl.urbaniak@gmail.com
https://ugent.academia.edu/RafalUrbaniak

Pawel Pawlowski
Centre for Logic and Philosophy of Science, Ghent University
haptism89@gmail.com
https://ugent.academia.edu/PawelPawlowski

Accepted and forthcoming in
INTRODUCTION TO FORMAL PHILOSOPHY
S. O. Hansson, V. Hendrick, and K. Esther Michelsen (eds), Springer, 2018.

Introduction Provability logics are, roughly speaking, modal logics meant to capture the formal principles of various provability operators (which apply to sentences) or predicates (which apply to sentence names). The first candidate for a provability logic was the modal logic **S4**, which contains as axioms all the substitutions of classical tautologies (in the language with $\Box$; throughout this survey when talking about instances or substitutions we'll mean instances and substitutions in the full language of the system under consideration), all substitutions of the schemata:

$$\begin{array}{ll} \text{(K)} & \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi) \\ \text{(M)} & \Box\varphi \to \varphi \\ \text{(4)} & \Box\varphi \to \Box\Box\varphi \end{array}$$

and is closed under two rules of inference: *modus ponens* (from $\vdash \varphi$ and $\vdash \varphi \to \psi$ infer $\vdash \psi$), and *necessitation* (Nec): if $\vdash \varphi$, then $\vdash \Box\varphi$.

The principles of **S4** seem sensible when $\Box\varphi$ is read as 'it is provable that φ': if an implication and its antecedent are provable, then so is its consequent, whatever is provable should be true, and if something is provable, we can prove that it is (by simply displaying the proof). The system was used in 1933 by Gödel to interpret intuitionistic propositional calculus (which is closely related to reasoning about provability). Alas, **S4** turned out to be inadequate as a tool for modeling the behavior of *formal provability predicate* within axiomatic arithmetic, mostly due to the fact that (M), also (in the context of provability logics) called *local reflection*, while intuitively plausible, cannot be provable in a consistent sufficiently strong axiomatic arithmetic for the formal provability predicate of that arithmetic. Let us elaborate.

Let's fix our attention on the standard first-order axiomatic arithmetic called *Peano Arithmetic* (**PA**). With this system in the background, instead of talking about an arithmetical formula φ, we can use a coding to represent it by some natural number, denoted by $\ulcorner\varphi\urcorner$. Once we've done this, there is (a standard way to construct) an arithmetical formula $\mathtt{Prov}_{\mathbf{PA}}(x)$ true exactly about the codes of those formulas, which are provable in **PA**. This is the formal provability predicate of **PA**.

One crucial property of this predicate is stated by *Löb's Theorem*, according to which for any arithmetical φ, if $\mathbf{PA} \vdash \texttt{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner) \rightarrow \varphi$, then already $\mathbf{PA} \vdash \varphi$. This means that reflection can hold only for those sentences which are already theorems of **PA**, and not universally for all sentences of arithmetic, and so **S4** cannot be the logic of formal provability predicate.

It turns out that another modal logic is the provability logic of formal arithmetical provability — it's the *Gödel-Löb logic* **GL**. Its axioms are all the substitutions of classical tautologies, all the substitutions of (K), all the substitutions of:

$$\text{(Löb)} \quad \Box(\Box\varphi \rightarrow \varphi) \rightarrow \Box\varphi$$

and the rules are *modus ponens* and necessitation. Various modal logics similar to **GL** have been developed for various notions of provability related to the standard formal provability.

In the language of **GL** we can express claims such as 'p is provable', but we cannot express things such as 't is a proof of p' (that is, we cannot express *explicit provability statements*). The latter task can be achieved in the so-called *Logic of Proofs* (**LP**), whose language is much richer: it contains terms for proofs, ways of constructing complex terms for proofs, and a predicate '_ is a proof of _'. **LP** is an adequate logic of explicit provability. Various extensions of **LP** has been developed.

Somewhat independently of the research on the logic of the formal provability predicate, attempts have been made to develop a formal logic of informal mathematical provability, for which (M) holds. The challenge is to develop a sensible system which can be mixed with other parts of mathematics without running into inconsistency due to (Löb) or related reasons.

This survey discusses the developments described above in a bit more detail.

Contents

1 The beginnings

a. Modal logic S4

Formulas of *the language of a propositional modal logic* $\mathcal{L}_M$ are built from propositional variables $p_1, p_2, \ldots$, two propositional constants $\bot$ (contradiction) and $\top$ (logical truth), classical connectives $\neg, \wedge, \vee, \rightarrow, \equiv$, brackets, and unary modal connectives $\Box$ and $\Diamond$, in the standard manner. Sometimes, without loss of generality, we'll treat $\mathcal{L}_M$ as containing only a single classical connective and a single modal operator — this will shorten some definitions, and is enough to make all the other connectives definable. Given a formal language (not necessarily $\mathcal{L}_M$, the context will make the range of meta-variables clear on each occasion), we'll use lower case Greek letters $\varphi, \psi, \chi, \ldots$ as meta-variables for formulas of that language (sometimes, we'll also use σ as a metavariable for an arithmetical *sentence*).

A *normal modal logic* contains as axioms all the substitutions of formulas of $\mathcal{L}_M$ for propositional variables in classical tautologies, all substitutions (in $\mathcal{L}_M$) of the schema:

$$\text{(K)} \qquad \Box(\varphi \rightarrow \psi) \rightarrow (\Box\varphi \rightarrow \Box\psi)$$

and is closed under two rules of inference: *modus ponens* (from $\vdash \varphi$ and $\vdash \varphi \rightarrow \psi$ infer $\vdash \psi$) and *necessitation* (Nec): if $\vdash \varphi$, then $\vdash \Box\varphi$. The weakest normal modal logic is called **K**, all other normal logics are its extensions.

The standard semantics of $\mathcal{L}_M$ involves *relational models* (also called *Kripke models).* A *frame* $\mathcal{F}$ is a tuple $\langle W, R\rangle$, where W is a non-empty set of possible worlds (or simply nodes, if you're not too much into bloated terminology) and R is a binary relation on W ('is a possible world from the perspective of'), often called an *accessibility relation*. A model $\mathbf{M}$ over $\mathcal{F} = \langle W, R\rangle$ is a triple $\langle W, R, \Vdash\rangle$, where $\Vdash$ is a *forcing* (or *satisfaction*) relation between W and the formulas of $\mathcal{L}_M$ (think about it as 'being true in'), satisfying the following conditions for any $w \in W$ and any $\varphi, \psi \in \mathcal{L}_M$:

$$\begin{array}{ll} w \nVdash \bot & w \Vdash \top \\ w \Vdash (\varphi \rightarrow \psi) & \text{iff } w \nVdash \varphi \text{ or } w \Vdash \psi \\ w \Vdash \Box\varphi & \text{iff for all } w' \in W, \text{ if } wRw', \text{ then } w' \Vdash \varphi \end{array}$$

It turns out that the class of formulas forced in every node in every frame is exactly the class of theorems of **K**. Sound and complete semantics for various other normal modal logics is obtained by putting further conditions on R.

One modal logic that will be of particular interest for us is **S4**, which (in one of the formulations) is obtained from **K** by adding as axioms all the instances of the following schemata:

$$\text{(M)} \qquad \Box\varphi \rightarrow \varphi$$

$$\text{(4)} \qquad \Box\varphi \rightarrow \Box\Box\varphi$$

(M) is sometimes called (T), but in what follows we'll often use $\mathbf{T}$ as a variable for an axiomatic theory, so to avoid confusion, we'll stick to (M). **S4** is sound and complete with respect to frames in which the accessibility relation is reflexive ($\forall w \in W\ wRw$) and transitive ($\forall w_1, w_2, w_3 \ \in W(w_1Rw_2 \wedge w_2Rw_3 \rightarrow w_1Rw_3)$).

Modal connectives of various modal systems admit various interpretations. $\Box$ can be interpreted as logical necessity, metaphysical necessity, physical necessity, moral obligation, knowledge, etc.[1] Different modal systems are taken to capture principles essential for these various notions. In what follows, we'll be concerned with the reading on which $\Box\varphi$ means 'it is provable that φ' (this reading will need further specifications, as it will turn out). Now the question is: which modal logic captures adequately the formal principles that hold for this reading?

Prima facie, **S4** seems like a decent candidate. (K) holds, because the consequent of a provable implication whose antecedent is provable is also provable. (M) holds, because whatever is provable is true. (4) holds, because if φ is provable, then by producing a proof of φ, by the same token, you are proving that it is provable (necessitation is reliable for pretty much the same reason). But are these considerations satisfactory? Not completely. First of all, we still don't know if there aren't any principles that hold for provability but are not provable in **S4**, because the argument so far was about the soundness of **S4** with respect to our intuitions about provability, not about completeness. Secondly, the argument is somewhat handwavy — it would be good to have a more precise explication of the notion of provability involved. Thirdly, even with such an explication in hand, we have to double-check if all principles of **S4** hold with respect to this explication. Things will turn out to be more complicated than one might initially expect.

b. Intuitionism and S4

S4 was first proposed as a logic of provability in the context of Brouwer's intuitionistic logic, which, very roughly speaking, results from replacing the notion of truth with that of constructive provability. The intuitionstic logic was formalized by Heyting (1930) as Intuitionistic Propositional Calculus (**IPC**) (see also Troelstra and van Dalen, 1988). On the intuitionistic approach, a mathematical claim is true just in case it has a proof, and false just in case there is a proof that it leads to contradiction. This idea inspired Heyting and Kolmogorov (Heyting, 1931, 1934; Kolmogorov, 1932) to introduce the so-called *Brouver-Heyting-Kolmogorov* (BHK) semantics, which identifies truth with provability, falsehood with refutability, and further specifies:

> A proof of $\varphi \wedge \psi$ consists of a proof of φ and a proof of ψ.
> A proof of $\varphi \vee \psi$ is provided by giving either a proof of φ or a proof of ψ.
> A proof of $\varphi \rightarrow \psi$ is a construction of proofs of ψ from proofs of φ.
> $\bot$ has no proof and $\neg\varphi$ means $\varphi \rightarrow \bot$.

Gödel (1933) attempted to formalize the BHK semantics. He put forward **S4** as the logic of classical provability. Then, he suggested a translation t from the non-modal language of intuitionistic logic into $\mathcal{L}_M$ by taking a non-modal formula and putting a box in front of each of its subformulas (in fact, this translation is already mentioned in (Orlov, 1928)). Gödel proved that if $\mathbf{IPC} \vdash \varphi$, then $\mathbf{S4} \vdash t(\varphi)$. The implication in the opposite direction has been later on proved by McKinsey and Tarski (1948). Thus,

[1] Notice however that different interpretations might make different principles plausible. For instance, (M) is not too convincing in the deontic reading, for unfortunately, not all that should be the case indeed is the case.

IPC, in a sense, can be taken to be about the classical provability, if, indeed, **S4** is the logic of classical provability (there are other modal logics into which IPC can be translated). Alas, an explicit provability semantics of $\Box$ in **S4** was missing, and so, the picture wasn't quite complete. One natural candidate for the interpretation of $\Box$ was a formal provability predicate in a standard axiomatized mathematical theory, to which we will now turn.

c. Arithmetical provability predicate

Considerations of formal provability predicate (or predicates) are usually developed in the context of an axiomatic arithmetic. This is the case for various reasons: via Gödel coding, instead of expressions, we can talk about numbers, standard arithmetical theories are usually strong enough to include a sufficiently reach theory of syntax (*modulo* coding), and arithmetic in general is a field where many results are already known and can be borrowed and applied to syntax.

For the sake of simplicity, we'll focus on one fairly standard axiomatic arithmetic: Peano Arithmetic (**PA**), although many results apply to other arithmetical theories, including some weaker ones (see for example Hájek and Pudlak, 1993, for details). The language of **PA**, $\mathcal{L}_{\mathbf{PA}}$, is a first-order language with identity and a few specific symbols: $0, S, \times$ and $+$ (in the standard model of arithmetic $\mathbb{N}$ interpreted as referring to the number zero, the successor function, multiplication, and addition, respectively). For any number m, the *standard numeral* for m has the form $\underbrace{S \ldots S}_{m} 0$ and is abbreviated by $\overline{m}$.

The specific axioms of $\mathbf{PA}$ consist of:

(PA 1) $$\forall x\ (0 \neq Sx)$$

(PA 2) $$\forall x, y\ (Sx = Sy \rightarrow x = y)$$

(PA 3) $$\forall x\ (x + 0 = x)$$

(PA 4) $$\forall x, y\ (x + Sy = S(x + y))$$

(PA 5) $$\forall x\ (x \times 0 = 0)$$

(PA 6) $$\forall x, y\ (x \times Sy = (x \times y) + x)$$

and all the instances of the induction schema:

(PA Ind) $$\varphi(0) \wedge \forall x\ (\varphi(x) \rightarrow \varphi(S(x))) \rightarrow \forall x\ \varphi(x)$$

Formulas of $\mathcal{L}_{\mathbf{PA}}$ can be classified according to their logical complexity. If t is a term not containing x, $\forall x \leq t\ \varphi(x)$ and $\exists x \leq t\ \varphi(x)$ abbreviate $\forall x\ (x \leq t \rightarrow \varphi(x))$ and $\exists x\ (x \leq t \wedge \varphi(x))$ respectively. Such occurrences of quantifiers are called *bounded*, and formulas whose all quantifiers are bounded are called Δ_0-formulas. The hierarchy proceeds in two "layers", that of Π_n and that of Σ_n formulas. $\Pi_0 = \Sigma_0 = \Delta_0$. Σ_{n+1}-formulas are of the form $\exists x_1, \ldots, x_k\ \varphi(x_1, \ldots, x_k)$, where $\varphi(x_1, \ldots, x_k)$ is Π_n. Π_{n+1}-formulas are of the form $\forall x_1, \ldots, x_k\ \varphi(x_1, \ldots, x_k)$, where $\varphi(x_1, \ldots, x_k)$ is Σ_n. Every formula of $\mathcal{L}_{\mathbf{PA}}$ is logically equivalent to a Σ_n formula and to a Π_m formula, for some n and m (and there always exist the least such n and m).

The class of Σ_1 formulas is of particular interest, because it turns out that a function is recursively enumerable (see Smith, 2007, for a nice introduction to the topic) just in case it is Σ_1-definable. This result, for instance, makes sure that an axiomatic system which is strong enough to handle Σ_1-sentences (in a sense to be specified) is strong enough to properly handle computable functions, including those related to syntactic manipulations, and so is strong enough to prove things about syntax of a formal language within it.

We say that an arithmetical theory $\mathbf{T}$ is Σ_1-*sound* just in case for any Σ_1-formula φ, if $\mathbf{T} \vdash \varphi$, then $\mathbb{N} \models \varphi$ (that is, φ is true in the standard model of arithmetic). The dual notion is that of Σ_1-*completeness*. $\mathbf{T}$ is Σ_1-complete just in case for any sentence $\varphi \in \Sigma_1$, if $\mathbb{N} \models \varphi$, then $\mathbf{T} \vdash \varphi$.

Fact 1.1. $\mathbf{PA}$ *is* Σ_1*-complete.*

There are various ways of *coding syntax*, effectively mapping syntactic objects, such as expressions, formulas, sentences and sequences thereof to natural numbers, so that each syntactic object τ of $\mathcal{L}_{\mathbf{PA}}$ is represented by its Gödel code $\ulcorner \tau \urcorner$. The details are unimportant here, so let's just focus on one of them and work with it (again, see Smith, 2007, for an accessible introduction).

Consider now any theory $\mathbf{T}$ in $\mathcal{L}_{\mathbf{PA}}$ extending $\mathbf{PA}$. It is said to be *elementary presented* just in case there is an arithmetical Δ_0-formula $\mathtt{Ax}_{\mathbf{T}}(x)$ true of a natural number just in case it is a code of an axiom of $\mathbf{T}$. Such a formula can be further used in a fairly standard way to construct a Δ_0 arithmetical formula $\mathtt{Prf}_{\mathbf{T}}(y, x)$ which is the standard binary proof predicate of $\mathbf{T}$ such that it is true of natural numbers m and n just in case m is the code of a sequence of formulas which is a proof of the formula whose code is n (the details of the construction are inessential here). Moreover:

(Binumeration) If in the standard model $\mathtt{Prf}_{\mathbf{T}}(m, n)$, then $\mathbf{PA} \vdash \mathtt{Prf}_{\mathbf{T}}(\overline{m}, \overline{n})$
If in the standard model $\neg\mathtt{Prf}_{\mathbf{T}}(m, n)$, then $\mathbf{PA} \vdash \neg\mathtt{Prf}_{\mathbf{T}}(\overline{m}, \overline{n})$

$\mathtt{Prf}_{\mathbf{T}}(y, x)$ can be further used to define the so-called *standard provability predicate* (since we won't be talking about non-standard provability predicates, we'll simply talk about provability predicates, assuming they're standard) and the *consistency statement*:

$$\mathtt{Prov}_{\mathbf{T}}(x) := \exists y\ \mathtt{Prf}_{\mathbf{T}}(y, x)$$
$$\mathtt{Con}(\mathbf{T}) := \neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner \bot \urcorner)$$

$\mathtt{Prov}_{\mathbf{T}}(y)$ is obtained from a Δ_0 formula by preceding it with an existential quantifier, and so, it is a Σ_1-formula. Therefore, by Σ_1-completeness, the first half of (Binumeration) holds for it (and the second one fails, for somewhat more complicated reasons):

If in the standard model $\mathtt{Prov}_{\mathbf{T}}(n)$ is true, then $\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\overline{n})$

Note however, that even though the second half of (Binumeration) fails, $\mathtt{Prov}_{\mathbf{T}}(x)$ succeeds at *defining* provability, in the sense that $\mathtt{Prov}_{\mathbf{T}}(\overline{\ulcorner \varphi \urcorner})$ is true in the standard model of arithmetic just in case in fact $\mathbf{T} \vdash \varphi$ (by the way, from now on we'll skip using the bar above numbers coding of formulas, assuming it is normally there, that is, that in the formulas we'll mention, numerals of codes of formulas are standard).

Still assuming $\mathbf{T}$ is elementary presented, $\mathtt{Prov}_{\mathbf{T}}(x)$ satisfies the following so-called *Hilbert-Bernays conditions* (Hilbert and Bernays, 1939; Löb, 1955) for any arithmetical formulas φ, ψ:

(HB1) $$\mathbf{T} \vdash \varphi \text{ iff } \mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$$

(HB2) $$\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \rightarrow \psi \urcorner) \rightarrow (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner))$$

(HB3) $$\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \mathtt{Prov}_{\mathbf{T}}(\ulcorner \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \urcorner)$$

In particular, the provability predicate of $\mathbf{T}$ can be taken to be that of $\mathbf{PA}$ itself. Also, keep in mind, that most of the results apply to certain theories weaker than $\mathbf{PA}$ and to elementary presented theories extending **PA**, either of which we usually chose to ignore for the sake of simplicity.

Another important piece of the puzzle will be Gödel's *incompleteness theorems*, which we include here in a somewhat modernized statement:

Theorem 1.2. *If an elementary presented theory* $\mathbf{T}$ *extends* $\mathbf{PA}$ *and is consistent, then there is a sentence* $G \in \mathcal{L}_{\mathbf{PA}}$ *such that* $\mathbf{T} \nvdash G$ *and* $\mathbf{T} \nvdash \neg G$*. Moreover,* $\mathbf{T} \nvdash \mathtt{Con}(T)$.

Incompleteness follows from a more general result (which have been stated by Carnap (1934); see (Gaifman, 2006) for a deeper historical discussion):

Lemma 1.3 (Diagonal Lemma)**.** *For any formula* $\varphi(x) \in \mathcal{L}_{\mathbf{PA}}$ *there is a sentence* $\lambda \in \mathcal{L}_{\mathbf{PA}}$ *such that*

$$\mathbf{PA} \vdash \lambda \equiv \varphi(\ulcorner \lambda \urcorner)$$

The Diagonal Lemma, when we take $\varphi(x)$ to be $\neg\mathtt{Prov}_{\mathbf{PA}}(x)$, entails the existence of a sentence that can be used in the incompleteness proof, which provably satisfies the condition:

$$G \equiv \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner G \urcorner)$$

Such a G is independent of $\mathbf{PA}$. The result generalizes: if a theory satisfies certain requirements and is consistent, its Gödel sentence is independent of it.

Quite a few years later Henkin (1952) asked a related question: what happens, however, with sentences such as:

(Henkin) $$H \equiv \mathtt{Prov}_{\mathbf{T}}(\ulcorner H \urcorner)?$$

The question was soon answered by Löb (1955):

Theorem 1.4 (Löb)**.** *If the Diagonal Lemma applies to* $\mathbf{T}$*, and the provability predicate of a theory* $\mathbf{T}$ *satisfies (his formulation of) the Hilbert Bernays conditions (HB1-3),* $\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \varphi$ *if and only if* $\mathbf{T} \vdash \varphi$.

For a given sentence φ, the formula '$\mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \rightarrow \varphi$' is called *reflection for* φ *(over **T**)*, and Löb's theorem says that reflection is provable in $\mathbf{T}$ for all and only theorems of $\mathbf{T}$. The theorem can be obtained fairly easily from the Diagonal Lemma applied to $\mathtt{Prov}_{\mathbf{T}}(x) \to \varphi$. For if the Diagonal Lemma applies to $\mathbf{T}$ (it is enough that $\mathbf{T}$ extends **PA**), the Lemma entails the existence of a ψ such that:

(L) $$\mathbf{T} \vdash \psi \equiv (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi)$$

The rest of the reasoning is propositional.

1	$\mathbf{T} \vdash (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi) \to \psi$	(L), Classical logic
2	$\mathbf{T} \vdash \psi \to (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi)$	(L), Classical logic
3	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \to (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi)\urcorner)$	(HB1), 2
4	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi)\urcorner)$	(HB2), 3
5	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to (\mathtt{Prov}_{\mathbf{T}}(\ulcorner \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner)\urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner))$	(HB2), 4
6	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner)\urcorner)$	(HB3)
7	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$	5, 6
8	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi$	Assumption
9	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner) \to \varphi$	7, 8
10	$\mathbf{T} \vdash \psi$	Classical logic, 1, 8
11	$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \psi \urcorner)$	(HB1), 10
12	$\mathbf{T} \vdash \varphi$	9, 11

d. The inadequacy of S4 with respect to formal provability

Coming back to the question of whether $\Box$ of **S4** can be sensibly interpreted as the formal provability predicate: what happens when we take $\Box\varphi$ to mean $\mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$? As it turns out, things fall apart quite quickly. For the sake of simplicity we'll take the case where $\mathbf{T} = \mathbf{PA}$, but the point generalizes to consistent recursively axiomatizable extensions of **PA**.

Since $\mathbf{S4} \vdash \Box\varphi \to \varphi$ for any φ, the interpretation would require that for all $\varphi \in \mathcal{L}_{\mathbf{PA}}$, $\mathbf{PA} \vdash \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \varphi \urcorner) \to \varphi$. But this, jointly with Löb's theorem, would entail that for any $\varphi \in \mathcal{L}_{\mathbf{PA}}$, $\mathbf{PA} \vdash \varphi$. So, if **PA** is consistent, **S4** is not the logic of the formal provability predicate of **PA**.

There is a somewhat different way to notice the inadequacy of **S4** in this context, already brought up by Gödel (1933). The formula expressing $\mathtt{Con}(PA)$ is $\neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner \bot \urcorner)$, which is logically equivalent to $\mathtt{Prov}_{\mathbf{PA}}(\ulcorner \bot \urcorner) \to \bot$. At the modal level, this is just an axiom of **S4**, since $\Box\bot \to \bot$ falls under schema (M). Thus, if **S4** was adequate, we would have $\mathbf{PA} \vdash \mathtt{Con}(PA)$, which would contradict Gödel's second incompleteness theorem. Moreover, necessitation would yield $\Box(\Box\bot \to \bot)$, and so in **S4** we would be able to derive the claim that the consistency claim is derivable, which again, contradicts Gödel's second incompleteness theorem.

At this stage, at least two questions remain open. What is the right modal logic of formal provability? What's the right provability semantics for S4?

2 Gödel-Löb modal logic (GL)

a. Axiomatizing GL

The inadequacy of **S4** was mainly due to Löb's theorem. How to proceed to obtain a modal logic better fit to the formal provability interpretation?

The first move results from noticing that it was (M) that was responsible for the trouble. Consequently, (M) has to be dropped. Another step is to notice that a formalized version of Löb's theorem can be proved in any elementary presented $\mathbf{T}$ extending **PA**, so that we have:

$$\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{T}}(\ulcorner \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner) \to \varphi \urcorner) \to \mathtt{Prov}_{\mathbf{T}}(\ulcorner \varphi \urcorner)$$

So, our modal logic of provability should validate the corresponding modal principle:

(Löb) $$\Box(\Box\varphi \to \varphi) \to \Box\varphi$$

GL (from Gödel-Löb) is a modal system resulting from these moves. Its axioms are all the substitutions of classical tautologies (in the language of $\mathcal{L}_M$), all the substitutions of (K), all the substitutions of (Löb), and the rules are *modus ponens* and necessitation.

Note that while (Nec) is a rule of **GL**, we cannot have $\varphi \to \Box\varphi$ as an axiom schema. While (Nec) is well-motivated (it says, in the intented interpretation, that any theorem is provably provable), the implication would say that anything *true* is provable, and that is far from obvious. In the arithmetical setting, we already have the formalized version of the second incompleteness theorem:

$$\mathbf{PA} \vdash \mathtt{Con}(\mathbf{PA}) \to \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner \mathtt{Con}(\mathbf{PA}) \urcorner)$$

so if we also had:

$$\mathbf{PA} \vdash \mathtt{Con}(\mathbf{PA}) \to \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \mathtt{Con}(\mathbf{PA}) \urcorner)$$

it would follow that $\mathbf{PA} \vdash \neg\mathtt{Con}(\mathbf{PA})$.

Now, is **GL** at least sound with respect to the formal provability interpretation? Well, the necessitation rule is the modal version of (HB1) and (K) is the modal version of (HB2). We can also prove in GL the modal version of (HB3), that is, $\mathbf{GL} \vdash (4)$, and so it can also be dropped when moving from **S4** to **GL**.

To get a better grasp of proofs in GL, let's see what the proof of this fact looks like. Before we give the proof we need two introductory steps. For one thing, since we have (Nec) and (K), we can easily move from $\mathbf{GL} \vdash (\varphi \to \psi)$ or from $\mathbf{GL} \vdash \Box(\varphi \to \psi)$ to $\mathbf{GL} \vdash \Box\varphi \to \Box\psi$. In what follows we'll make such moves without hesitation, sometimes calling them (Distr) — since they basically consist in distributing $\Box$ over material implication. If you're not convinced, the official argument starts with $\vdash \varphi \to \psi$. Use (Nec) to obtain $\vdash \Box(\varphi \to \psi)$. Then (K) tells you $\vdash \Box(\varphi \to \psi) \to (\Box\varphi \to \Box\psi)$, and so by detachment $\vdash \Box\varphi \to \Box\psi$. It's just too repetitive to go through these moves every time.

For another, we'll need this fairly straightforward fact:

Fact 2.1. $\mathbf{GL} \vdash \Box(\varphi \wedge \psi) \equiv (\Box\varphi \wedge \Box\psi)$

Proof. Reason within **GL**. From left to right:

1.	$\varphi \wedge \psi \to \varphi$	Tautology
2.	$\varphi \wedge \psi \to \psi$	Tautology
3.	$\Box(\varphi \wedge \psi) \to \Box\varphi$	(Distr), 1
4.	$\Box(\varphi \wedge \psi) \to \Box\psi$	(Distr), 2
5.	$\Box(\varphi \wedge \psi) \to (\Box\varphi \wedge \Box\psi)$	Classical logic, 3, 4

From right to left:

1.	$\varphi \to (\psi \to (\varphi \wedge \psi))$	Tautology
2.	$\Box\varphi \to (\Box\psi \to \Box(\varphi \wedge \psi))$	(Distr), 1
3.	$(\Box\varphi \wedge \Box\psi) \to \Box(\varphi \wedge \psi)$	Classical logic, 2

□

Observe also that Fact 2.1 entails (left to right, together with conjunction elimination):

$$\mathbf{GL} \vdash \Box(\Box\varphi \wedge \varphi) \to \Box\Box\varphi \tag{2.2}$$

Finally we have:

Fact 2.3. $\mathbf{GL} \vdash (4)$*, that is* $\mathbf{GL} \vdash \Box\varphi \to \Box\Box\varphi$.

Proof. Again, let's reason within GL.

1.	$\varphi \to ((\psi \wedge \chi) \to (\chi \wedge \varphi))$	Tautology
2.	$\varphi \to ((\Box\Box\varphi \wedge \Box\varphi) \to (\Box\varphi \wedge \varphi))$	Substitution, 1
3.	$\varphi \to (\Box(\Box\varphi \wedge \varphi) \to (\Box\varphi \wedge \varphi))$	Fact 2.1, 2
4.	$\Box\varphi \to \Box(\Box(\Box\varphi \wedge \varphi) \to (\Box\varphi \wedge \varphi))$	(Distr), 3
5.	$\Box(\Box(\Box\varphi \wedge \varphi) \to (\Box\varphi \wedge \varphi)) \to \Box(\Box\varphi \wedge \varphi)$	(Löb)
6.	$\Box\varphi \to \Box(\Box\varphi \wedge \varphi)$	Classical logic, 4, 5
7.	$\Box\varphi \to \Box\Box\varphi$	(2.2), 6

□

We've shown that (4) is derivable in **GL**. But since (M) was the source of the problems, we also need to make sure it is not a derivable theorem schema for **GL**. Simply dropping it from the axiom schemata is not enough.

Fact 2.4. *If* $\mathbf{GL} \nvdash \bot$*, it is not the case that for any* φ*,* $\mathbf{GL} \vdash \Box\varphi \to \varphi$.

Proof. Suppose the opposite holds. Then we have $\mathbf{GL} \vdash \Box\bot \to \bot$, and so we can reason within GL:

1.	$\Box\bot \to \bot$	Assumption
2.	$\Box(\Box\bot \to \bot)$	(Nec), 1
3.	$\Box(\Box\bot \to \bot) \to \Box\bot$	(Löb)
4.	$\Box\bot$	Detachment , 2, 3
5.	$\bot$	Detachment, 1, 4

□

The assumption of consistency of **GL** above is explicit not because we have any serious doubts about it. It's rather that when it is explicitly stated, the unprovability of reflection can be easily proven in a few steps, as we've just seen. Proof of a similar claim without this assumption is slightly more convoluted.

Fact 2.5. $\mathbf{GL} \nvdash \bot$ *and* $\mathbf{GL} \nvdash \Box p \to p$*.*

Proof. The general structure of the argument is this. We show that all theorems of **GL** have a certain property, which $\bot$ and $\Box p \to p$ don't have, and so $\bot$ and $\Box p \to p$ are not theorems of **GL**. The property is: *being a classical propositional tautology under the following translation*. So now we need to define a translation t from $\mathcal{L}_M$ into the classical propositional language, which translates all theorems of **GL** into classical tautologies, but at the same time translates $\bot$ and $\Box p \to p$ into formulas whose negations are classically satisfiable. Let's start with the translation:

$$\begin{aligned} t(\bot) &= \bot \\ t(p) &= p \text{ (for all propositional variables)} \\ t(\varphi \to \psi) &= (\varphi)^\star \to (\psi)^\star \\ t(\Box\varphi) &= \top \end{aligned}$$

Clearly:

1. If φ is a substitution of a classical tautology, $t(\varphi)$ is a tautology. This is because the translation effectively is a substitution, and it gives a formula in the classical propositional language, in which all substitutions of tautologies are classical tautologies.

2. t(K) is $\top \to (\top \to \top)$, which is a classical tautology.

3. t(Löb) is $\top \to \top$, which also is a tautology.

We handled the axioms of **GL**, making sure their translations are classical tautologies. Now we need to take care of the inference rules.

4. Consider *modus ponens* (arguments for any classical propositional rule are pretty much the same). One can still apply *modus ponens* to $t(\varphi)$, $t(\varphi \to \psi) = (t(\varphi) \to t(\psi))$. So if $\mathbf{GL} \vdash \varphi, \mathbf{GL} \vdash \varphi \to \psi$, we know that $\mathbf{GL} \vdash \psi$, and that the following are tautologies: $t(\varphi)$, $t(\varphi) \to t(\psi)$, and $t(\psi)$.

5. What about necessitation? Say $\mathbf{GL} \vdash \varphi$ so that also $\mathbf{GL} \vdash \Box\varphi$. Quite trivially $t(\Box\varphi) = \top$, which is a tautology.

Together, points 1-5 show that all theorems of **GL** translate into classical tautologies. Finally, we have to show that the translations of the formulas that we're interested in aren't tautologies.

6. $t(\Box p \to p) = \top \to p$, which is not a tautology.

7. $t(\bot) = \bot$, which also isn't a tautology.

Points 6-7 mean that these formulas are not theorems of **GL**, which completes the proof. □

b. Another way towards GL: K4LR

Another modal logic that might come to mind when one thinks of $\Box$ as provability is **K4LR**. Just as **GL**, it allows necessitation and classical consequence (for the modal language), and just as **GL** it has (K) as an axiom schema. But it keeps (4), drops (Löb), and admits the following *Löb's rule* (LR) instead:

(LR) $\qquad$ If $\vdash (\Box\varphi \to \varphi)$, infer $\vdash \Box\varphi$

It turns out that **GL** and **K4LR** have the same theorems.

Fact 2.6. *If* **K4LR** $\vdash \varphi$*, then* **GL** $\vdash \varphi$.

Proof. We need to check that **GL** proves the axioms of **K4LR** and that it is closed under its rules. As for the axioms, (K) is shared, and Fact 2.3 shows that **GL** $\vdash$(4). As for the rules (Nec) is shared, and we only need to show that **GL** is closed under (LR). This can be shown by the following reasoning within **GL**:

1.	$\Box\varphi \to \varphi$	Assumption (as a GL-theorem)
2.	$\Box(\Box\varphi \to \varphi)$	(Nec), 1
3.	$\Box(\Box\varphi \to \varphi) \to \Box\varphi$	(Löb)
4.	$\Box\varphi$	MP, 2, 3
5.	φ	MP, 1, 4

□

Implication in the opposite direction also holds.

Fact 2.7. *If* **GL** $\vdash \varphi$*, then* **K4LR** $\vdash \varphi$.

Proof. (K) and (Nec) and classical logic in the background are shared. The only thing that needs to be shown is **K4LR** $\vdash$(Löb), that is that (LR) in the context of **K4LR** is strong enough to give us the formula corresponding to the rule.

To see that this is not immediately obvious, note that, in principle, rules are weaker than corresponding implications, because they apply to theorems only. For instance, (Nec) is sensible because it says that any theorem is necessary, but the formula $p \to \Box p$ is not an axiom of any sensible standard modal logic, for it says, roughly, that *any truth is necessary*. Let's prove (Löb) within **K4LR**.

1.	$\Box(\Box\varphi \to \varphi) \to (\Box\Box\varphi \to \Box\varphi)$	(K)
2.	$\Box(\Box\varphi \to \varphi) \to \Box\Box(\Box\varphi \to \varphi)$	(4)
3.	$\Box[\Box(\Box\varphi \to \varphi) \to \Box\varphi] \to [\Box\Box(\Box\varphi \to \varphi) \to \Box\Box\varphi]$	(K)
4.	$\Box[\Box(\Box\varphi \to \varphi) \to \Box\varphi] \to [\Box(\Box\varphi \to \varphi) \to \Box\Box\varphi]$	PL, 1, 3
5.	$\Box[\Box(\Box\varphi \to \varphi) \to \Box\varphi] \to [\Box(\Box\varphi \to \varphi) \to \Box\varphi]$	PL, 2, 4
6.	$\Box(\Box\varphi \to \varphi) \to \Box\varphi$	(LR), 5

Line 1 applies (K) to the antecedent of (Löb). In line 2 we use (4) to modalize the antecedent of (Löb) even deeper. Line 3 applies axiom (K) to distribute necessity over the antecedent and the consequent of (Löb). By classical logic, lines 2 and 3 allow us to replace $\Box\Box(\Box\varphi \to \varphi)$ with $\Box(\Box\varphi \to \varphi)$ in the antecedent of the consequent of the formula in line 3. In line 5, thanks to line 1 we could remove one box in the last consequent of the formula in line 4. Now we notice that line 5 is just the premise for an application of (LR) and we apply this rule. □

c. GL vs. logical necessity

What difference does it make to read $\Box$ as 'it is provable in the system' rather than as 'it is logically necessary'? Well, (K), (4) and necessitation intuitively speaking hold for both readings. But there are some important differences.

One thing, already mentioned, is that **GL** $\nvdash \Box p \rightarrow p$, while all suitable candidates for a modal logic of logical necessity (the main one being S5) prove $\Box p \rightarrow p$. And rightly so, for it seems intuitively true that whatever is logically necessary holds.

What about (Löb)? It obviously is a theorem of **GL** (yes, we're sloppy about the distinction between formula and schemata, but this shouldn't cause any misunderstanding). But if we read $\Box$ as logical necessity, it is somewhat difficult to sort out our modal intuitions about (Löb). Notice that our intuitions about necessity validate reflection for all formulas. Among them is:

$$\Box\bot \rightarrow \bot$$

This seems right: if a contradiction is necessary, it is true (hopefully, it isn't). Also, the implication seems to be a logical truth itself, and as such should also be necessary. So we have:

$$\Box(\Box\bot \rightarrow \bot)$$

as an intuitive truth when $\Box$ is read as logical necessity. At the same time, we (with a few notable exceptions, such as Graham Priest) don't think there are true contradictions – so, *a fortiori*, we don't think there are necessary contradictions. This gives us:

$$\neg\Box\bot$$

Put those two things together, and we easily have:

$$\neg[\Box(\Box\bot \rightarrow \bot) \rightarrow \Box\bot]$$

So, when we read $\Box$ as logical necessity, we have an intuitively convincing formula which is the negation of an instance of (Löb)! And in general, it seems false that just because all necessary sentences are true, all sentences are true.

Another way to see why (Löb) is problematic when $\Box$ is read as necessity is this. Substitute $\neg p$ and apply a few trivial classical moves and the fact that $\Diamond\varphi$ is equivalent to $\neg\Box\neg\varphi$:

$$\begin{aligned}\Box(\Box\neg p \rightarrow \neg p) &\rightarrow \Box\neg p\\ \Box\neg(p \wedge \Box\neg p) &\rightarrow \Box\neg p\\ \neg\Box\neg p &\rightarrow \neg\Box\neg(p \wedge \Box\neg p)\\ \Diamond p &\rightarrow \Diamond(p \wedge \Box\neg p)\end{aligned}$$

But the last formula says that if p is possible then it is possible that p and yet p is necessarily false. This surely isn't an intuitively convincing principle of logical necessity.

d. GL and deontic modalities

Multiple sources, when introducing **GL**, mention (Smiley, 1963) — a paper titled "The Logical Basis of Ethics" as the source in which the Löb's theorem stated as a modal formula first occured. However, to our knowledge, none of these sources actually explains what it was doing there. For instance, in (Verbrugge, 2016) all that we can read about it is:

> Ironically, the first time that the formalized version of Löb's theorem was stated as the modal principle [...] was in a paper by Smiley in 1963 about the logical basis of ethics, which did not consider arithmetic at all.

We'd like to be more specific, and so a short digression about how (Löb) entered the stage follows.

Smiley is developing the ideas from (Anderson, 1956), where an attempt is made to define a deontic modality O (*it is obligatory that*) in terms of an alethic one (*it is necessary that*). Anderson's basic idea is to define:

$$O\varphi =_{df} \Box(\neg\varphi \to S)$$

where S is an unspecified constant expressing the claim that some sanction is applied. Given this analysis, a modal logic of obligation is obtained via a translation from a modal logic of necessity (as far as Anderson's system is concerned, the resulting system is **S4** with O instead of $\Box$).

(Smiley, 1963) contains an interesting discussion of a philosophical concern raised by Nowell-Smith and Lemmon (1960) as to whether S is supposed to contain a deontic aspect, or is it supposed to be purely factive, and different difficulties arising in these two cases, but let's put these issues aside. Smiley's point, however, is that a contrapositive reformulation of Anderson's account makes the connection between the left-hand side and the right-hand side more intuitive. Instead of meeting a sanction if ϕ is not performed, Smiley talks about being a consequence of a moral code, so that:

$$O\varphi \equiv \Box(T \to \varphi)$$

where T is a moral code. Now, Smiley argues, we can ask what inferential principles hold for O no matter which particular T is chosen.

An important difference in Smiley's approach, however, is that the above equivalence is not definitional, and so the respective formulas aren't replaceable in all contexts. This is like treating the equivalence as an assumption (to which (Nec) cannot be applied) rather than as an axiom schema. This reading, he claims, validates (K), application of (Nec) to tautologous formulas, but not axioms (4) or (M). Accordingly, he conjectures that the right modal logic will comprise exactly these.

Further on, Smiley discusses another reading, on which $O\varphi$ holds if φ follows from the totality of obligatory propositions. But then, there is no single sentence expressing this totality, and so instead of using T as a sentential symbol, Smiley considers using it as an operator, so that $T\varphi$ means that φ belongs to this totality. While one might object that it is not clear what philosophical progress can be made by analysing being an obligatory sentence in terms of following from the totality of obligatory sentences (especially as, presumably, all obligatory sentences that follow from the totality of obligatory sentences are already in it), we can still ask what formal properties O thus defined would have. Smiley's reply is that given that such operators aren't treated in any modal logic, we should turn to arithmetic and the arithmetical provability predicate, which, arguably, might have similar formal properties as 'being a consequence of a moral code' (so the claim in (Verbrugge, 2016) that the paper didn't consider arithmetic at all is a bit hasty). A this point, Smiley observes that on this arithmetical reading, all tautologous formulas are theorems, (K) and (4) hold, and so do *modus ponens* and (Nec). Then, Smiley mentions Löb's theorem saying:

> ... and Kripke has pointed out to me that this proof can itself be arithmetised to provide a proof of the formula $O(OA \to A) \to OA$.
>
> (Smiley, 1963, 244)

So, indeed, Smiley does mention the formula in the context of ethics.

Now, just a few words about what *doesn't* happen in the paper. Smiley doesn't explain how (Löb) would be understood if the modality is interpret as a deontic modality, doesn't discuss any philosophical motivations for (Löb) in this interpretation (independent of the behavior of provability in arithmetic), and doesn't propose (Löb) as an additional axiom of a modal logic of obligation.

Come to think of it, (Löb) in the deontic reading, doesn't seem too plausible. On one hand, it should be the case that *whatever* should be the case happens (i.e. obligations should be obeyed). On the other hand, it seems unintuitive that just because of that, simply anything *whatsoever* should be the case.

e. GL and formal provability

We know **S4** turned out inadequate with respect to formal provability predicate. **GL** does a much better job. To elaborate, we first need to explain the relation between $\mathcal{L}_M$ and $\mathcal{L}_{\mathbf{PA}}$ that will underlie what follows.

A mapping from propositional variables of $\mathcal{L}_M$ to the set of sentences of $\mathcal{L}_{\mathbf{PA}}$ is called an *arithmetical realization*. In a sense, an arithmetical realization tells us which variables are to be interpreted as which sentences of arithmetic. Given an elementary presented theory **T**, any arithmetical realization r can be extended to a **T**-interpretation $r_{\mathbf{T}}(\varphi)$ of a modal formula, by the following conditions:

$$r_{\mathbf{T}}(\bot) = \bot \quad r_{\mathbf{T}}(\top) = \top$$
$$r_{\mathbf{T}}(p) = r(p) \text{ for any variable } p$$
$$r_{\mathbf{T}}(\varphi \to \psi) = r_{\mathbf{T}}(\varphi) \to r_{\mathbf{T}}(\psi)$$
$$r_{\mathbf{T}}(\Box\varphi) = \mathtt{Prov}_{\mathbf{T}}(\ulcorner r_{\mathbf{T}}(\varphi) \urcorner)$$

If you worry that $\mathcal{L}_{\mathbf{PA}}$ doesn't really contain $\bot$ and $\top$, feel free to replace them with any $\mathcal{L}_{\mathbf{PA}}$-formulas that are, respectively, refutable and provable by pure logic. Let's call the set of all possible **T**-interpretations of $\varphi \in \mathcal{L}_M$ (under all possible realizations) $\varphi_{\mathbf{T}}$.

Given the correlation between the axioms and rules of **GL** and the Hilbert-Bernay's conditions and Löb's theorem, adequacy of **GL** at least in one direction is clear:

Fact 2.8. $\mathbf{GL}$ *is sound with respect to the arithmetical interpretation, that is:*

$$\textit{If } \mathbf{GL} \vdash \varphi, \textit{ then } \mathbf{PA} \vdash \varphi_{\mathbf{T}}.$$

(where by $\mathbf{PA} \vdash \varphi_{\mathbf{T}}$ *we mean that* $\mathbf{PA}$ *proves all the members of* $\varphi_{\mathbf{T}}$*).*

In fact, implication in the opposite direction also holds, provided that $\mathbf{T}$ is Σ_1-sound, so that the claim can be strengthened to equivalence (Solovay, 1976):

Theorem 2.9 (Solovay Completeness)**.** *If* $\mathbf{T}$ *is* Σ_1*-sound, then for any* $\varphi \in \mathcal{L}_M$*:*

$$\mathbf{GL} \vdash \varphi \ \textit{ if and only if } \ \mathbf{T} \vdash \varphi_{\mathbf{T}}.$$

This shows that given a sensible arithmetical theory, those principles of its formal provability predicate that are provable in arithmetic are adequately axiomatized by **GL**. The proof lies beyond the scope of this survey, but the general strategy can be quickly described. Assume $\mathbf{GL} \nvdash \varphi$. Then, by the results to be described in section 2.f. (feel free to read this passage again after reading that section) there is a finite transitive and reversely well-founded model such that for some w in it, $w \nVdash \psi$. Since the set of worlds in the model W is finite, we can safely identify W with an initial segment of natural numbers $= \{1, 2, \ldots, n\}$ with $w = 1$ and $1Ri$ just in case $1 < i \leq n$. The tricky part now, the part for which Solovay is deservedly famous, is using this arithmetical counterpart of W to construct an interpretation such that the arithmetical theory fails to prove the arithmetical interpretation of φ.

f. Relational semantics for GL

We have drawn a connection between **GL** and the formal provability predicate. What about relational semantics for **GL**, though? As it turns out (Segerberg, 1971), there is a natural class of relational models with respect to which **GL** is sound and complete.

Theorem 2.10. $\mathbf{GL}$ *is sound and complete with respect to the class of finite frames in which* R *is transitive and irreflexive.*

There is a somewhat different class of frames with respect to which **GL** is sound and complete. We say that the accessibility relation R is *reversely well-founded* in W just in case every non-empty subset X of W has an R-maximal element (that is, a $w \in X$ such that $\neg\exists w' \in W \ wRw'$).

Theorem 2.11. $\mathbf{GL}$ *is sound and complete with respect to transitive and reversely well-founded frames.*

Notice that there is a connection between these two. Any reversely well-founded R is irreflexive, and a transitive R on a finite set is reversely well-founded just in case it is irreflexive. The result can be strengthened:

Theorem 2.12. $\mathbf{GL}$ *is sound and complete with respect to finite transitive and reversely well-founded frames.*

Since the proof employs a construction that given a formula to be checked gives an upper limit on the finite size of models to be checked, the proof by the same token proves the decidability of **GL**.

The full proof of weak completeness (that is, the one that applies to theoremhood, read on for details) is beyond the scope of this survey. To give you a taste, however, we'll run the following interesting part of the argument to the effect that if (Löb) holds in a frame, its accessibility relation is reversely well-founded. We'll argue by contraposition, by showing that if a frame isn't reversely well-founded, there is a possible world in it and a forcing relation over it, such that (Löb) fails there.

So assume R is not reversely well-founded. This means there is a set $X \subseteq W$ such that the elements of X constitute an infinite chain $w_1 R w_2 R w_3 \ldots$. Take $\Vdash$ such that $w \Vdash p$ for all $w \in W \setminus X$ and $w' \Vdash \neg p$ for all $w' \in X$. Pick an arbitrary $w \in X$. Now we want to show that the antecedent of (Löb), $\Box(\Box p \rightarrow p)$, holds in w. This requires showing that $\Box p \rightarrow p$ holds in any world accessible from w. So assume wRv. We'll want to show $v \Vdash \Box p \rightarrow p$.

Either $v \in X$ or $v \notin X$. If the former, then v can access at least one world in the infinite chain. So for some $u \in X$, vRu. Since p is false in all elements of X we have $u \Vdash \not p$ and so $v \Vdash \Diamond\neg p$, that is $v \Vdash \neg\Box p$. But this classically entails $v \Vdash \Box p \rightarrow p$. If the latter, $v \Vdash p$, and classically $v \Vdash \Box p \rightarrow p$.

Either way, if wRv, $v \Vdash \Box p \rightarrow p$. Since our choice of v was arbitrary, and the only assumption was that wRv, this means that $w \Vdash \Box(\Box p \rightarrow p)$. This is the antecedent of (an instance) of (Löb). On the other hand, w is in a chain in X, and so it can access a world where p fails, and so $w \Vdash \neg\Box p$, which is the negation of (our instance of) (Löb).

Notice that the property of being conversely well-founded isn't first-order definable.[2] That is, there is no first-order formula containing the binary predicate letter R which holds in a model just in case R is conversely well-founded. For suppose there is such a formula ψ. Introduce infinitely many new constants $c_1, c_2, \ldots$. Consider the infinite set of formulas composed of ψ and $\{c_i R c_j \mid i < j\}$ (which jointly state the existence of an infinite chain). Each finite subset of that set is satisfiable in a model (take any finite conversely well-founded model, where there are more objects than constants under consideration). But then, by *compactness theorem for first-order classical languages* (which in one of its formulations says that if any finite subset of a set of first-order formulas is satisfiable, then so is the whole set), the whole set has a model. Among other things, this model makes ψ true (in which no new constants occur), and yet, R in it cannot be conversely well-founded, because it has to contain an infinite R-chain of objects corresponding to the new constants.

Compactness (to be elaborated on in section 2.g.), as used in the above argument, holds for first-order logic, and our argument was about first-order definability, so its use in this context is legitimate. This issue shouldn't be confused with the question of compactness of **GL**, because, as it will turn out, **GL** itself is not compact.

One remark: soundness and completeness in the above theorems is taken in the *weak* sense: ψ is valid in all finite, transitive and reversely well-founded frames just in case it is a theorem of **GL**. This sense is to be distinguished from strong soundness and

[2] See (Blackburn et al., 2001) for an extensive introduction to issues related to definability of properties of frames.

completeness. To introduce this notion we have to define **GL**-*derivability* first. We say that ψ is **GL**-derivable from (a finite or infinite) premise set Γ ($\Gamma \vdash_{\mathbf{GL}} \psi$) just in case there is a proof of ψ from the axioms of **GL** and formulas belonging to Γ by the rules of **GL**, provided (Nec) is applied only to theorems of **GL** (alternatively, iff there is a proof of ψ from *theorems of* **GL** and elements of Γ by means of *modus ponens* only; the latter formulation has the advantage that the deduction theorem to the effect that $\vdash \varphi \rightarrow \psi$ just in case $\varphi \vdash \psi$ applies to derivability thus defined). Now, the relevant strong completeness claim is that $\Gamma \vdash_{\mathbf{GL}} \psi$ just in case in every possible world in every finite, transitive and reversely well-founded model, if all elements of Γ are true in it, then so is ψ. Alas, the claim is false — strong completeness for **GL** fails, pretty much for the same reasons for which compactness fails for **GL**. We'll explain this in more detail soon.

Given the arithmetical soundness and completeness, relational semantics provides us with a handy tool for showing that a certain claim about provability is not provable in **PA**: to show this it is enough that its modal counterpart is not provable in **GL**, and given the relational semantics, to show this it is enough to construct a transitive reversely well-founded model making the claim false. A nice example (Boolos, 1993) is that of

$$\Box(\Box p \vee \Box\neg p) \rightarrow (\Box p \vee \Box\neg p)$$

Is it provable in **GL**? The answer is negative. Consider a model composed of three possible worlds a, b, c such that aRb, aRc, $c \Vdash \neg p$, $b \Vdash p$. Since b and c are blind worlds, $b, c \Vdash \Box p, \Box\neg p, \Box p \vee \Box\neg p$, and so $a \Vdash \Box(\Box p \vee \Box\neg p)$. Yet, a sees a world where p and a world where $\neg p$, and so $a \nVdash \Box p \vee \Box\neg p$.

This means there is an arithmetical sentence σ which can be assigned to p by a realization, such that the negation of the resulting **PA**-interpretation of the formula in question can be consistently added to it.

In other words, it is consistent with **PA** that it is provable that either σ is provable or refutable, but nevertheless σ is neither provable nor refutable. And this clearly is a nice little fact that will score you some extra points in a late night conversation with a stranger in a pub.

g. Compactness failure for GL

Compactness was already mentioned in section 2.f.. We say that a logic **L** is *compact* just in case for any set Γ of formulas of the appropriate formal language, if every finite subset of Γ has a model suitable for **L**, the whole Γ has such a model. In the case of **GL**, compactness would mean that for every set Γ of formulas of $\mathcal{L}_M$, if every finite subset of Γ has a transitive and conversely well-founded model, then so does Γ. Interestingly, compactness fails for **GL**.

To see why, first reflect on the meaning of $\Diamond$ in this context. $\Diamond\varphi$ is defined as $\neg\Box\neg\varphi$, and so while '$\Box\varphi$' is read as 'φ is provable', the intuitive reading of '$\Diamond\varphi$' is 'φ is not refutable' or 'the negation of φ is not provable'.

Now we can proceed with the reasoning. Take an infinite assembly of propositional variables $p_0, p_1, \ldots$, and the following infinite set of formulas:

$$C = \{\Diamond p_0\} \cup \{\Box(p_i \rightarrow \Diamond p_{i+1}) \mid i \in \mathbb{N}\}$$

Every finite subset of C has a transitive reversely well-founded model. There is no space for the general argument here, but to see why this is plausible consider the finite subset of C:

$$C_2 = \{\Diamond p_0, \Box(p_0 \to \Diamond p_1), \Box(p_1 \to \Diamond p_2)\}$$

Take (the transitive closure of) the frame composed of w, w_0, w_1, w_2 only, such that $wRw_0Rw_1Rw_2$. Whether propositional variables are true at w is irrelevant, for the other worlds take the forcing relation such that $w_i \Vdash p_i$ and $w_i \nVdash p_j$ if $i \neq j$. We'll argue that $w \Vdash C_2$.

- Since wRw_0 and $w_0 \Vdash p_0$, $w \Vdash \Diamond p_0$.
- $w_2 \nVdash p_0$, so $w_2 \Vdash p_0 \to \Diamond p_1$. For a similar reason, $w_1 \Vdash p_0 \to \Diamond p_1$. Moreover, since $w_1 \Vdash p_1$, $w_0 \Vdash \Diamond p_1$, and so $w_0 \Vdash p_0 \to \Diamond p_1$. This shows that $p_0 \to \Diamond p_1$ holds in all worlds accessible from w. So $w \Vdash \Box(p_0 \to \Diamond p_1)$.
- A perfectly analogous argument goes for $w \Vdash \Box(p_1 \to \Diamond p_2)$.

However, C doesn't have a transitive and conversely well-founded model. For suppose there is a model with a w such that $w \Vdash C$. Define:

$$X = \{v \mid wRv \wedge \exists i \ v \Vdash p_i\}$$

That is, collect all the possible worlds accessible from w where at least one p_i holds. We have $\Diamond p_0 \in C$, so $w \Vdash \Diamond p_0$, and X is non-empty, say $wRw_0, w_0 \Vdash p_0$. Since $w \Vdash \Box(p_0 \to \Diamond p_1)$, $w_0 \Vdash p_0 \to \Diamond p_1$. So $w_0 \Vdash \Diamond p_1$, and there is a $w_1 \in X$ such that $w_1 \neq w_0$ (R is irreflexive) such that w_0Rw_1 and $w_1 \Vdash p_1$. But $w \Vdash \Box(p_1 \to \Diamond p_2)$ and (by transitivity) wRw_1, and so $w_1 \Vdash p_1 \to \Diamond p_2$. Therefore $w_1 \Vdash \Diamond p_2$, and so w_1 has to see yet another member of X, etc. In short: X has to contain an infinite chain, which contradicts the assumption that R is conversely well-founded.

The example of C can be also used to explain why strong completeness fails for **GL**. We already know C has no transitive, conversely well-founded model. Another way to say this is that C semantically entails $\bot$ (with respect to this class of frames): $C \models \bot$. Yet, $\bot$ is not derivable from C, $C \nvdash \bot$ — for any proof from C could only use a finite number of premises from C, and we already know that no finite subset of C entails (and so, by soundness, proves) $\bot$.

The fact that compactness fails is partially responsible for why the semantic completeness for **GL** is a bit more tricky than one for a more usual modal logic. Normally, in the proof, one constructs a canonical model by taking infinite sets of consistent formulas as possible worlds; for **GL**, however there are syntactically consistent sets of formulas which nevertheless aren't semantically coherent.

h. Leterless sentences and the normal form theorem for GL

A *letterless sentence* of $\mathcal{L}_M$ is a formula built from the classical and modal connectives, devoid of propositional variables, and containing only $\bot$ among its atomic formulas. Instead of preceding φ with n boxes, we'll write $\Box^n\varphi$.

One of the reasons why letterless sentences are interesting is because some of them formally certain fairly natural statements. For instance, $\neg\texttt{Prov}_{\mathbf{T}}(\ulcorner\bot\urcorner)$ is (=formalizes) the consistency statement (hopefully true, but under standard conditions unprovable), $\mathtt{Prov}_{\mathbf{T}}(\ulcorner\neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner\bot\urcorner)\urcorner)$ is the provability of consistency (false), $\neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner\bot\urcorner) \rightarrow \neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner\neg\mathtt{Prov}_{\mathbf{T}}(\ulcorner\bot\urcorner)\urcorner)$ expresses the second incompleteness theorem, etc.

One of the interesting uses of $\mathbf{GL}$ (to which we will move soon) relate to the existence of a certain decision procedure, whose description employs the notion of a normal form. By a *normal form of a letterless sentence* φ we mean a truth-functional combination of sentences of the form $\Box^i\bot$.

Theorem 2.13 (Normal form theorem (Boolos))**.** *For any letterless formula $\varphi \in \mathcal{L}_M$, there is a normal form letterless ψ such that $\mathbf{GL} \vdash \varphi \equiv \psi$.*

The normal form theorem, while at first it might seem abstract, will come handy quite soon, see section 2.k..

i. ω-consistency

An arithmetical theory $\mathbf{T}$ is *ω-inconsistent* just in case there is a formula $\psi(x)$ in the language of $\mathbf{T}$ such that for each $n \in \mathbb{N}$ we have $\mathbf{T} \vdash \psi(\overline{n})$, and yet, we also have $\mathbf{T} \vdash \exists x\ \neg\psi(x)$. $\mathbf{T}$ is *ω-consistent* iff it is not ω-inconsistent.

ω-inconsistency doesn't entail inconsistency *simpliciter*. After all, $\mathbf{T}$ can have a non-standard model, where all the standard numbers have the property expressed by ψ, and yet, some non-standard number (not named by a numeral) is a witness to $\exists x\ \neg\psi(x)$ (see Kaye, 1991; Hedman, 2004; Kossak, 2006, for more details on non-standard models of arithmetic). By the same token, consistency doesn't entail ω-consistency either.

Now, in the standard interpretation of $\mathbf{GL}$, $\Box$ represents provability. What is represented by $\Diamond$? Well, by definition $\Diamond\varphi$ just in case $\neg\Box\neg\varphi$, and so $\Diamond\varphi$ holds just in case the negation of φ is not provable – that is, just in case φ is consistent (with the underlying axiomatic system of arithmetic). In this sense, $\mathbf{GL}$ can be thought of as a logic of consistency.

The question arises — what is the logic of ω-consistency? Can its propositional principles be axiomatized? The answer is easier than expected.

Let $\omega\mathtt{Con}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$ be the arithmetical formula expressing the ω-consistency of φ with an elementary presented theory $\mathbf{T}$. If we think of it as $\Diamond\varphi$, then $\Box\varphi$ corresponds to $\neg\omega\mathtt{Con}_{\mathbf{T}}(\ulcorner\neg\varphi\urcorner)$. Accordingly, let's modify the definition of realisation so that:

$$r_{\mathbf{T}}(\Box\varphi) = \neg\omega\mathtt{Con}_{\mathbf{T}}\ulcorner\neg r_{\mathbf{T}}(\varphi)\urcorner.$$

Theorem 2.14. *Given the above conventions, the set of always provable formulas is axiomatized by* $\mathbf{GL}$*, and the set of always true formulas is axiomatized by* $\mathbf{S}$ *(an extension of* $\mathbf{GL}$ *described in section 3.a.).*

j. Provability in analysis

Roughly speaking, analysis is second-order arithmetic, that is, arithmetic with second-order logic available, where the infinity of instances of the induction schema is replaced with a single induction axiom (see however Simpson, 2009, for more details and variety of higher-order systems):

$$\forall P\ [(P(0) \wedge \forall x\ (Px \rightarrow P(Sx)) \rightarrow \forall x\ Px]$$

What is the logic of provability of analysis? Again, no surprises: it is **GL**.

Now imagine we want to strengthen the system with the so-called *ω-rule* will allows to infer $\forall x\ \psi(x)$ from $\psi(\overline{n})$ for all $n \in \mathbb{N}$. Note: ψ is ω-inconsistent with $\mathbf{T}$ just in case $\neg\psi$ is derivable from $\mathbf{T}$ by *one* application of the ω-rule.

It turns out that this move doesn't change the underlying logic of provability. Still, **GL** is the modal logic of provability in analysis with the ω-rule.

k. Applications of GL

Solovay completeness allows us to use **GL** to make inferences about provability predicates of elementary presented theories. Let's call a sentence of $\mathcal{L}_{\mathbf{PA}}$ a *constant sentence of* **PA** if it belongs to the least set of formulas containing $\bot$ (if you don't like $\bot$ being explicitly in $\mathcal{L}_{\mathbf{PA}}$ take it to be $0 = 1$), closed under classical connectives, such that if ψ is a constant sentence, then so is $\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\psi\urcorner)$ (the notion generalizes to other theories). The notion was introduced by Harvey Friedman, who asked whether there is an effective decision procedure for evaluating the truth-value of constant sentences. The answer is positive and relies on Theorem 2.13. The procedure is this:

- Take the letterless $\varphi \in \mathcal{L}_{\mathbf{PA}}$ and find the letterless $\psi \in \mathcal{L}_M$ such that $r_{\mathbf{T}}(\psi) = \varphi$ (notice, for letterless sentences, the choice of r is irrelevant).
- Put ψ in the normal form.
- $\Box^i\bot$ has the same truth value as $\bot$, so delete all $\Box^i$ in front of $\bot$.
- We are left with a sentence in the non-modal language of propositional logic, devoid of propositional variables. Evaluate it. It is true just in case so is φ.

One nice general result about **GL** that has interesting consequences is De Jongh-Sambin fixed point theorem. To introduce it, some preliminaries are needed. A formula φ of $\mathcal{L}_M$ is said to be *modalized in the propositional variable* p just in case every occurrence of p in φ is within the scope of $\Box$ (this also applies to vacuous cases, so formulas devoid of p are also modalized in p). A formula is called a *p–formula* if it contains no variable other than p.

A formula φ is called a *fixed point* of a formula ψ with respect to variable p just in case φ contains only those sentence letters that occur in ψ, doesn't contain any occurrence of p, and:

$$\mathbf{GL} \vdash \Box(p \equiv \psi) \equiv \Box(p \equiv \varphi)$$

Theorem 2.15 (De Jongh-Sambin fixed point theorem)**.** *If ψ is modalized in p, there is a fixed point φ for ψ relative to p.*

This form of the theorem is useful for eliminating apparent self-reference from arithmetical sentences: finding their provably equivalent counterparts which do not contain self-reference. For instance, if ψ is $\neg\Box p$, the fixed point is $\neg\Box\bot$. So, by fixed point theorem:

$$\mathbf{GL} \vdash \Box(p \equiv \neg\Box p) \equiv \Box(p \equiv \neg\Box\bot)$$

By arithmetical soundness of **GL**, for any arithmetical sentence χ, we have $\mathbf{PA} \vdash \chi \equiv \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\chi\urcorner)$ just in case $\mathbf{PA} \vdash \chi \equiv \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner\bot\urcorner)$. So χ, equivalent to its own unprovability turns out to be also provably equivalent to the consistency statement. A few more examples. The fixed point of $\Box p$ is $\top$. So if we take χ provably equivalent to its own provability, the fixed point theorem tells us that we can equally well describe χ without self-reference, in the sense that:

$$\mathbf{PA} \vdash \chi \equiv \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \chi \urcorner) \text{ iff } \mathbf{PA} \vdash \chi \equiv \top$$

and similarly:

$$\mathbf{PA} \vdash \chi \equiv \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \neg\chi \urcorner) \text{ iff } \mathbf{PA} \vdash \chi \equiv \mathtt{Prov}_{\mathbf{PA}}(\bot)$$
$$\mathbf{PA} \vdash \chi \equiv \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner \neg\chi \urcorner) \text{ iff } \mathbf{PA} \vdash \chi \equiv \bot$$

The utility of the fixed point theorem might perhaps become more clear if we look at a somewhat different formulation. Let $\psi(p)$ be a formula containing p among propositional variables occurring in it.

Theorem 2.16 (De Jongh-Sambin, second formulation)**.** *For any $\psi(p) \in \mathcal{L}_M$ modalized in p, there is a formula $\varphi \in \mathcal{L}_M$ containing only variables from ψ, not containing p, such that:*

$$\mathbf{GL} \vdash \varphi \equiv \psi(\varphi)$$

Any fixed points of $\psi(p)$ are provably equivalent in $\mathbf{GL}$.

In a sense, fixed point theorem is the modal counterpart of the Diagonal Lemma. This formulation makes it clear why the theorem is called a fixed-point theorem. Generally, a fixed point of a function f is an argument such that $f(x) = x$, and φ is the fixed point of ψ because $\psi(\varphi) \equiv \varphi$.

Moreover, the proof is effective, in the sense that it provides a recipe for constructing appropriate fixed points. Some examples of formulas and their fixed point are:

Formula	Fixed Point
$\Box p$	$\top$
$\Box\neg p$	$\Box\bot$
$\neg\Box p$	$\neg\Box\bot$
$\neg\Box\neg p$	$\bot$
$q \wedge \Box p$	$q \wedge \Box q$

Consider the third formula. $\neg\Box p$ says that p isn't provable. Its fixed point is $\neg\Box\bot$, and so, by the fixed point theorem, we have:

$$\mathbf{GL} \vdash \neg\Box\bot \equiv \neg\Box(\neg\Box\bot)$$

But the arithmetical realization of $\neg\Box\bot$ for $\mathbf{T}$ is $\mathtt{Con}(\mathbf{T})$. So the above formula (from left to right):

$$\text{(G2)} \qquad \neg\Box\bot \to \neg\Box(\neg\Box\bot)$$

represents the formalized version of Gödel's second incompleteness theorem: if the theory is consistent, it doesn't prove its own consistency.

In fact, Gödel's second incompleteness can be fairly easy reached in **GL** without the full power of the fixed point theorem:

1	$\Box(\Box\bot \to \bot) \to \Box\bot$	(L)
2	$\neg\Box\bot \to \neg\Box(\Box\bot \to \bot)$	contraposition, 1
3	$\neg\Box\bot \to \neg\Box(\neg\Box\bot)$	def. of $\neg$, 2

Second incompleteness is about not being able to prove the consistency claim. This, however, can be strengthened to the undecidability of consistency, because in **GL** it is also possible to prove that if the inconsistency is not provable, then neither is the inconsistency claim:

1	$\Box\bot \to \bot$	(M)
2	$\Box(\Box\bot \to \bot)$	(Nec), 1
3	$\Box\Box\bot \to \Box\bot$	(Distr), 2
4	$\neg\Box\bot \to \neg\Box\Box\bot$	contraposition, 3

The modally formalized version of Gödel's first incompleteness theorem is:

(G1) $$\neg\Box\bot \to (\Box(p \equiv \neg\Box p) \to \neg\Box p)$$

It also can be proved within **GL** (we'll use "CL" to mark moves made by classical propositional logic):

1	$\Box p \to p$	(M)
2	$\Box\neg p \to \neg p$	(M)
3	$\Box p \wedge \Box\neg p \to \bot$	CL, 1, 2
4	$(\Box p \equiv \Box\neg p) \wedge \Box p \to \Box p \wedge \Box\neg p$	CL
5	$(\Box p \equiv \Box\neg p) \wedge \Box p \to \bot$	CL, 3, 4
6	$(\Box p \equiv \Box\neg p) \wedge \Box p \to \Box\bot$	CL, 5
7	$\Box(p \equiv \neg p) \to (\Box p \equiv \Box\neg p)$	(Distr)
8	$\Box(p \equiv \neg p) \wedge \Box p \to \Box\bot$	CL, 6,7
9	$\neg(\Box(p \equiv \neg p) \to \neg\Box p) \to \Box\bot$	CL, 8
10	$\neg\Box\bot \to (\Box(p \equiv \neg p) \to \neg\Box p)$	CL, 9

Another argument which is quite easy to run with **GL** at hand, is for the claim that no sentence consistent with **PA** can imply all reflection principles. For suppose that for any φ:

1	$\mathbf{GL} \vdash S \to (\Box\varphi \to \varphi)$	Assumption
2	$\mathbf{GL} \vdash S \to (\Box\neg S \to \neg S)$	Instance of 1
3	$\mathbf{GL} \vdash \Box\neg S \to \neg S$	CL, 2
4	$\mathbf{GL} \vdash \neg S$	(Löb), 3

The result means that in **PA** (and in any sensible **T** extending **PA**) we cannot finitely axiomatize reflection principles involving the provability predicate of any sensible $\mathbf{T}'$ extending **PA** (**PA** included).

Coming back to the fixed point theorem, observe that it doesn't hold for all formulas of $\mathcal{L}_M$. For instance, a fixed point of p (or $\neg p$) itself would be a letterless sentence S_p such that $\mathbf{GL} \vdash S_p \equiv p$ (or $\mathbf{GL} \vdash S_{\neg p} \equiv \neg p$), and it can be easily proven by induction on formula length that there is no such a letterless sentence.

3 Close kins of GL

a. Modal logic S

What happens, however, when instead of asking about those principles that are provable in the background arithmetic, we ask about those principles which are true in the standard model?

Modal system **S** is defined as the closure of **GL** together with (M) under *modus ponens* and substitution. Notice: (Nec) is inadmissible, apart from the job it does in generating the theorems of **GL**, included in **S**. Otherwise we could argue:

1.	$\Box\bot \to \bot$	(M)
2.	$\Box(\Box\bot \to \bot)$	(Nec), 1
3.	$\Box\bot$	(Löb), 2
4.	$\bot$	CL, 1, 3

The failure of (Nec) means that **S** is not a normal modal logic. Interestingly, despite the failure of (Nec) in **S**, **S** validates the inference from $\mathbf{S} \vdash \varphi$ to $\mathbf{S} \vdash \Diamond\varphi$, which is not validated by **GL**. For reflection for $\neg\varphi$ gives $\mathbf{S} \vdash \Box\neg\varphi \to \neg\varphi$, contraposition gives $\mathbf{S} \vdash \varphi \to \neg\Box\neg\varphi$. The result then follows by the definition of $\Diamond$ and modus ponens.

Now, let $S(\varphi)$ be:

$$(\Box\varphi_1 \to \varphi_1) \wedge \cdots \wedge (\Box\varphi_k \to \varphi_k)$$

where $\Box\varphi_1, \ldots, \Box\varphi_k$ are all subformulas of φ of the form $\Box\chi$. The following holds:

Theorem 3.1 (II Solovay Completeness)**.** *If* $\mathbf{T}$ *is sound (that is, for any* $\varphi \in \mathcal{L}_{\mathbf{PA}}$*, if* $\mathbf{T} \vdash \varphi$*, then* $\mathbb{N} \models \varphi$*), the following conditions are equivalent for any* $\psi \in \mathcal{L}_M$*:*

$$\mathbf{S} \vdash \psi$$
$$\mathbf{GL} \vdash S(\psi) \to \psi$$
$$\mathbb{N} \models \psi_{\mathbf{T}}$$

In a sense, Theorem 3.1 tells us that the only claims always true but not always provable are those which are required to make reflection hold.

Since **S** isn't a normal modal logic, it doesn't have straightforward relational models. A semantics for **S** in terms of the so-called *tail models* have been given by Visser (1984).

b. Strong provability and Grzegorczyk's Grz

Consider exending a realization r to the *Grzegorczyk* ***T****-interpretation* $r^G_{\mathbf{T}}$, which differs from **T**-interpretation in what it does with the modal operator:

$$r^G_{\mathbf{T}}(\Box\varphi) = r^G_{\mathbf{T}}(\varphi) \wedge \mathtt{Prov}_{\mathbf{T}}(\ulcorner r^G_{\mathbf{T}}(\varphi)\urcorner)$$

Thus, while the standard intepretation reads the box as *provable*, Grzegorczyk interpretation reads it as *true and provable* (Grzegorczyk, 1967). The arithmetically complete logic of strong provability (Boolos, 1993) is **Grz**, which is obtained from **S4** by adding:

(Grz) $$\Box(\Box(\psi \to \Box\psi) \to \psi) \to \psi$$

There is an interesting connection between most of the axioms of **Grz** and the properties of strong provability provable in a somewhat weaker system **K4** (that is, **GL** without (Löb)) (Smoryński, 2004). Define $[s]\varphi$ as $\Box\varphi \wedge \varphi$. Then, if **K4** $\vdash \varphi$, then also **K4** $\vdash [s]\varphi$. **K4** proves (all instances of) the following:

$$
\begin{aligned}
{[s]}\varphi \wedge [s](\varphi \to \psi) &\to [s]\psi \\
{[s]}\varphi \to [s][s]\varphi, \quad [s][s]\varphi &\to [s]\varphi \\
{[s]}\varphi &\to \varphi
\end{aligned}
$$

Note however, that if we replace $\Box$ in (Grz) with $[s]$, the result is not a **K4** theorem schema (it is invalidated by a single-world model, where ψ is false in the only possible world).

By the way, strong provability can be used in obtaining **GL** in yet another manner. For suppose you want to enrich **K4** with something that does the job of the diagonal lemma at the arithmetical level. To do this is, we need to say that if something is the consequence of a diagonalization, it should be a theorem. One way to capture this intuition is to add the *Diagonalization Rule*:

(DR) $$\text{If } \vdash [s](p \equiv \varphi(p)) \to \psi, \text{ then } \vdash \psi.$$

De Jongh has proven that **GL** is closed under (DR), and Smoryński has shown that **K4**+(DR) coincides with **GL**.

Grz is another example of a logic of provability into which a translation of inuitionistic **IPC** can be given. Extensions of **Grz** to the context of the logic of proofs have been further studied in (Nogina, 1994, 1996).

c. Provability of Σ_1-sentences (GLV and GLSV)

Σ_1 sentences have the specific property that for any such sentence σ

$$\mathbf{PA} \vdash \sigma \to \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \sigma \urcorner).$$

Accordingly, the modal logics of the provability of Σ_1 sentences has been characterized by Visser by first taking **GLV** to be an extension of **GL** with all formulas of the form $p \to \Box p$ (it's crucial that p is a propositional variable, because Σ_1 aren't closed under arbitrary Boolean combinations). The rules of **GLV** are the same as those of **GL** — *modus ponens* and (Nec). Then, **GLSV** has as axioms all the theorems of **GLV** and all instances of reflection, and its only rule of inference is *modus ponens*.

A realization r is a Σ_1-*realization* if for any propositional variable p, $r(p)$ is a Σ_1-sentence. Call a relational model a GLV-model just in case it is finite, irreflexive, transitive and such that for all $w, v \in W$ and all propositional variables p:

$$wRv, w \Vdash p \Rightarrow v \Vdash p$$

(the last condition means that accessiblility preserves the satisfaction of propositional variables).

Theorem 3.2 ((Visser, 1997, 2008)). **GLV** $\vdash \psi$ *just in case* ψ *is valid in all GLV-models just in case for all* Σ_1*-realization* r, $\mathbf{PA} \vdash r^{PA}(\psi)$. $\mathbf{GLSV} \vdash \psi$ *iff for all* Σ_1 *realizations* r, $r^{PA}(\psi)$ *is true in the standard model.*

4 The logic of proofs LP

a. Motivations

One of the reasons why thinking about provability is tricky, especially in the context of first-order theories, is that a universal quantifier is involved. Given that first-order arithmetical theories have non-standard models which contain non-standard numbers, this leads to certain troubles. In general, if we take a model M of an arithmetical theory, it might be the case that $\exists x\ \varphi(x)$ holds in that model, with no $\varphi(\overline{n})$ holding for $n \in \mathbb{N}$, because the existential formula has a non-standard witness.

In particular, this applies to $\mathtt{Prov}_{\mathbf{T}}(\ulcorner\varphi\urcorner)$, which in fact means $\exists x\ \mathtt{Prf}_{\mathbf{T}}(x, \ulcorner\varphi\urcorner)$. The problem is, this formula doesn't entail that for some $n \in \mathbb{N}$, $\mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner)$. This feature results in certain disparities in the behavior of $\mathtt{Prov}_{\mathbf{T}}(y)$ and $\mathtt{Prf}_{\mathbf{T}}(x, y)$.

The case where this is particularly visible is that of reflection. *Explicit reflection* has the form $\mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner) \to \varphi$. All instances of explicit reflection are provable in the underlying arithmetical theory $\mathbf{T}$ (satisfying our standard conditions). For either $\mathbb{N} \models \mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner)$ or $\mathbb{N} \not\models \mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner)$. In the former case, then indeed there is a proof of φ in $\mathbf{T}$, and since $\mathbf{T} \vdash \varphi$, by classical logic, $\mathbf{T} \vdash \mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner) \to \varphi$. If, on the other hand, it's not the case that $\mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner)$, then (since it's a Δ_0 formula, and we assumed $\mathbf{T}$ to be sufficiently strong) $\mathbf{T} \vdash \neg\mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner)$, and again, by classical logic, $\mathbf{T} \vdash \mathtt{Prf}_{\mathbf{T}}(\overline{n}, \ulcorner\varphi\urcorner) \to \varphi$. Either way, an arithmetical theory satisfying the standard strength requirements proves explicit reflection, for any $n \in \mathbb{N}$ and any $\varphi \in \mathcal{L}_{\mathbf{PA}}$.

In contrast, due to Löb's theorem, given a consistent and sufficiently strong $\mathbf{T}$, *local reflection* for $\mathtt{Prov}_{\mathbf{T}}(x)$:

$$\mathtt{Prov}_{\mathbf{T}}(\ulcorner\varphi\urcorner) \to \varphi$$

is provable only for those formulas, which are theorems of $\mathbf{T}$. Indeed, at the level of **GL**, due to (Löb), one cannot consistently add reflection $\Box\varphi \to \varphi$, for otherwise, the reasoning already described in section 3a. can be used to derive contradiction.

Clearly, $\mathcal{L}_M$ lacks the resources to represent explicit reflection, because $\Box$ represents provability, and $\Box\varphi \to \varphi$ can be used to represent local reflection only. A more expressive language to achieve that goal has been devised to underlie the logic of proofs, **LP** (Artemov, 1994, 1998, 2001).

b. The language and axioms of LP

The pure language of the logic of proofs extends the non-modal propositional language with new symbols:

- *proof variables* $(x, y, z, \dots)$ and *proof constants* $(a, b, c, \dots)$,
- three proof operation symbols: *binary application* $(\cdot)$, *binary union* (+) and *unary proof checker* (!),

- *is a proof of* symbol (:).

A *proof polynomial* is either a proof variable, or a proof constant, or is built from proof polynomials by means of $\cdot$, $+$, or !. Binary application intuitively corresponds to *modus ponens*, in the sense that if t is a proof of $\varphi \to \psi$, and s is a proof of φ, then $t \cdot s$ is a proof of ψ. The union of two proofs t and s, $t + s$ proves anything that either t or s does. The proof checker operation checks whether a given proof t of φ is correct, and if it is, it yields a proof that t proves φ.

The *is a proof of* symbol is used to construct atomic formulas from proof polynomials and formulas. If t is a proof polynomial and φ is a formula, $t{:}\varphi$ is a formula saying that t *is a proof of* φ, or simply t *proves* φ.[3]

One rule of **LP** is *modus ponens*. The axioms of **LP** (on top of classical propositional logic) are:

(Application) $t{:}(\varphi \to \psi) \to (s{:}\varphi \to (t \cdot s){:}\psi)$

(Reflection) $t{:}\varphi \to \varphi$

(Proof checker) $t{:}\varphi \to !t{:}(t{:}\varphi)$

(Sum) $s{:}\varphi \to (s + t){:}\varphi, t{:}\varphi \to (s + t){:}\varphi$

Another rule of **LP**, allows to introduce $c{:}\varphi$ as a theorem, whenever φ is an axiom, where c is a new constant in a given proof. In this context, the **LP**-counterpart of (Nec) is a derivable rule.

Fact 4.1. *If* **LP** $\vdash \varphi$*, then for some proof polynomial* p*,* **LP** $\vdash p{:}\varphi$*.*

Notice that **LP** is not a normal modal logic: we can't simply treat $t{:}$ as we would treat $\Box$ in a normal modal logic. For instance, (K) for $t{:}$ fails, as this is not generally the case:

$$t{:}(p \to q) \to (t{:}p \to t{:}q)$$

c. Properties of LP

LP is decidable (Mkrtychev, 1997). It is also sound and complete with respect to provability interpretation in **PA** (where proof polynomials are mapped to appropriate proof codes). Quite some time ago we seemed to have left **S4** behind, as arithmetically inadequate. One of the nice features of **LP** is that it brings **S4** back to the table. A *forgetful projection* of an **LP**-formula φ is a modal formula resulting from replacing all the occurences of $t{:}(\varphi)$ with $\Box(\varphi)$.

Theorem 4.2. *(Artemov, 1995) The forgetful projection of* **LP** *is* S4*.*

Possible worlds semantics for **LP** (requiring the extension of the standard framework with the so-called *evidence function*) has been developed by Mkrtychev (1997) and Fitting (2003).

[3] A short historical remark: Gödel suggested the use of explicit proofs to interpret **S4** in a public lecture in Vienna, without describing the logic. The content of the lecture has been published in 1995, and the logic of proofs was formulated independently before that publication.

LP itself doesn't allow to express (and *a fortiori*) prove mixed statements about explicit proofs and provability, which nevertheless seem of indepedent interest. For instance, in the provability semantics the following *explicit-implicit principle* is valid:

$$\neg(t{:}\varphi) \to \Box\neg(t{:}\varphi).$$

d. Mixed logic of explicit proofs and provability (B)

Such caims can be proven in a mixed logic of explicit proofs and provability, **B** (Artemov, 1994). The axioms are that of **GL** enriched with:

$$\begin{aligned} t{:}\varphi &\to \varphi \\ t{:}\varphi &\to \Box(t{:}\varphi) \\ \neg(t{:}\varphi) &\to \Box\neg(t{:}\varphi) \end{aligned}$$

and the *Rule of reflection*, which allows to infer $\mathbf{B} \vdash \varphi$ from $\mathbf{B} \vdash \Box\varphi$. Artemov proved also the following:

Theorem 4.3. $\mathbf{B}$ *is sound and complete with respect to the semantics of proofs and provability in* $\mathbf{PA}$.

For a survey of further studies and properties of **LP** and its extensions, see (Artemov, 2007) and other developments in (Nogina, 1994; Yavorskaya, 2001; Yavorsky, 2001).

5 Formal logics of informal provability

a. Motivations

Informal provability is closely related to what mathematicians do when they prove theorems, rather than to formal provability in an axiomatic system (Sundholm, 1998). A sentence is informally provable if it is provable by any commonly accepted mathematical means. According to the proponents of *the standard view* there is no important difference between formal and informal proofs. Any informal proof, they say, is just a sloppy and incomplete version of a fully formalized proof in an appropriate formal theory. Thus, informal provability reduces to formal provability within some axiomatic system, usually to some version of set theory.

Yet, some people disagree with the above picture (Myhill, 1960; Horsten, 1998; Leitgeb, 2009; Marfori, 2010):

- It is not clear which axiomatic system we should choose to represent informal provability. It seems that the informal notion of provability is unified whereas in different formal systems different theorems are provable.
- It is not clear how to convert an informal proof into a formal one.
- It is not clear whether we should associate each informal proof with exactly one formal proof or with some abstraction class of formal representations of informal proofs, and if yes, how such a class is to be identified.

- It is not clear whether the conversion to formal proofs preserves identity laws for informal proofs. It may be the case, that two substantially different informal proofs are associated with exactly the same formal representation.
- Formal proofs are stated in a fully formalized language. Informal proofs on the other hand are stated in the natural language expanded with additional mathematical vocabulary.
- The role of axioms is different in proofs of these two kinds. In formal proofs axioms are simply one of the syntactically admissible ways of extending a given proof. In an informal proof, axioms partially or fully explicate the meaning of the notions involved.
- The justification of subsequent steps is of a different nature. In formal proofs it's purely syntactical. In informal proofs, mathematicians often refer to semantical notions such as truth-preservation or mathematical intuition.
- The reflection schema, which says that if something is (informally) provable then it is true, is intuitively compelling for informal provability. Yet, as already discussed, for any sensible notion of formal provability, we cannot have it.

Since Gödel, however, there is an agreement as to which principles are intuitively correct for informal provability: those are the principles of **S4**. So, if we were to produce an axiomatization of those principles, which intuitively hold for informal provability, the validity of all the instances of the reflection schema is crucial. For formal provability, by Löb's theorem, we know that the reflection schema is only provable for theorems – and there is no independent philosophical motivation for this restriction to be imposed on informal provability.

One way out might be to strengthen the underlying axiomatic system by brute force by adding all the instances of the reflection schema. One thing to observe, however is that even a small amount of reflection schema turns out to be arithmetically strong:

Fact 5.1. *Let* $\mathbf{T}$ *be a theory consisting of* $\mathbf{PA}$ *and all the instances of the reflection schema for* $\mathtt{Prov}_{\mathbf{PA}}(x)$ *restricted to* Π_1 *formulas. Then* $\mathbf{T} \vdash \mathtt{Con}(\mathbf{PA})$.

Proof. Let φ be $\forall x\ x \neq x$. It is a Π_1 pure logical contradiction. Let's abbreviate it as $\bot$. By the assumption, $\mathbf{T} \vdash \mathtt{Prov}_{\mathbf{PA}}(\ulcorner \bot \urcorner) \to \bot$. By classical logic, $\mathbf{T} \vdash \neg\bot$. So we have $\mathbf{T} \vdash \neg\mathtt{Prov}_{\mathbf{PA}}(\ulcorner \bot \urcorner)$. □

But say we're not worried about intuitively obvious claims about provability turning out to be arithmetically strong. Another observation is that as far as formal provability is concerned, we can only consistently add reflection for the *old* formal provability predicate, thus obtaining a *new* formal theory with *new* formal provability predicate, for which reflection still fails to be provable.

Indeed, what we can consistently do still falls short of accepting reflection for the theory that we are working within. Suppose we extend $\mathcal{L}_{\mathbf{PA}}$ with an additional primitive symbol, a new provability predicate P, for which we add to our background arithmetical theory $\mathbf{T}$ (extending **PA**) all instances of the Hilbert-Bernays conditions and all the instances of the reflection schema, thus obtaining a new theory **TP**. We run into the following problem:

Fact 5.2. *No* $\mathbf{TP}$ *satisfying the conditions below is consistent.*

$$\mathbf{TP} \vdash P(\ulcorner\varphi\urcorner) \rightarrow \varphi$$
$$\textit{If } \mathbf{TP} \vdash \varphi, \ \textit{then } \mathbf{TP} \vdash P(\ulcorner\varphi\urcorner)$$
$$\textit{(HB1-3) for } P \textit{ hold.}$$

This and related results motivated various attempts to develop a formal logic of informal provability (which formally captures inferential principles intuitively valid for informal provability, most notably reflection) while avoiding such pitfalls. The main idea is that instead of constructing a formal provability predicate within arithmetic, one develops a logic of informal provability by introducing a new symbol for provability and considering various axioms and rules that might apply to it. We'll now take a look at the main candidates, which come in two flavors. The first group treats informal provability as an operator not as a predicate, thus blocking those inferential moves with are available for predicates, but not for operators, and thus avoiding contradiction at the cost of limited expressivity. The second group of theories treats informal provability as a predicate, but limit the scope the Hilbert-Bernays conditions for the new provability predicate. At the end of this section we'll also take a look at two stray dogs which don't really fit into any of these groups.

b. Epistemic arithmetic (EA)

Historically, the first theory of informal provability is Shapiro's *Epistemic Arithmetic* (**EA**) presented in (Shapiro, 1985) and developed by Goodman (1984) and Flagg and Friedman (1986). The idea here is to extend the standard arithmetical language $\mathcal{L}_{\mathbf{PA}}$ to $\mathcal{L}_K$ by adding a unary operator K that applies to formulas. The underlying arithmetical theory is **PA**, and the behavior of K is characterized by the following rules:

KI If $\Gamma \vdash \varphi$ and every element of Γ is epistemic, then $\Gamma \vdash K(\varphi)$
KE $K(\varphi) \vdash \varphi$

where a formula φ is ontic iff it does not contain any occurrences of the operator K and is epistemic iff it has the form $K(\psi)$ for some formula ψ. So EA has all axioms of $\mathbf{PA}$ and the above two rules for K. Note, the above rules imply **S4** principles for K.

Unfortunately, the internal logic of **EA** (that is, what in **EA** is provably provable) is quite a weak theory – in a sense, it is an elementary extension of intuitionistic Heyting Arithmetic (**HA**). Define a translation V from $\mathcal{L}_{HA}$, the language of **HA**, into $\mathcal{L}_K$. We use $\bar{\varphi}$ to indicate that φ belongs to $\mathcal{L}_{HA}$ as follows:

1. For atomic formulas: $V(\bar{\varphi}) = K(\bar{\varphi})$,
2. $V(\overline{\varphi \wedge \psi}) = K(V(\overline{\varphi})) \wedge K(V(\overline{\psi}))$,
3. $V(\overline{\varphi \vee \psi}) = K(V(\overline{\varphi})) \vee K(V(\overline{\psi}))$,
4. $V(\overline{\varphi \rightarrow \psi}) = K(K(V(\overline{\varphi})) \rightarrow K(V(\overline{\psi})))$,
5. $V(\overline{\varphi \equiv \psi}) = K(K(V(\overline{\varphi})) \equiv K(V(\overline{\psi})))$,

6. $V(\overline{\neg\varphi}) = K(\neg K(V(\overline{\varphi})))$,
7. $V(\overline{\forall x\ \varphi(x)}) = K(\forall x\ V(\overline{\varphi(x)}))$,
8. $V(\overline{\exists x\ \varphi(x)}) = \exists x\ KV(\overline{\varphi(x)})$.

Just for the sake of simplicity we will write φ instead of $\bar{\varphi}$ whenever it does not lead to confusion. The above translation is sound and complete in the following sense:

Theorem 5.3. *For every $\varphi \in \mathcal{L}_{HA}$, if $\mathbf{HA} \vdash \varphi$, then $\mathbf{EA} \vdash V(\varphi)$.*

Theorem 5.4 (Goodman 1984)**.** *For every $\varphi \in \mathcal{L}_{HA}$, if $\mathbf{EA} \vdash V(\varphi)$, then $\mathbf{HA} \vdash \varphi$.*

EA, however, does have some interesting properties — we'll mention only two of them. The *numerical existence property* is that for any formula φ, if $\mathbf{EA} \vdash \exists x\ K\varphi(x)$ then for some natural number n, $\mathbf{EA} \vdash K\varphi(n)$. The *disjunction property* is that if $\mathbf{EA} \vdash K(\varphi \vee \psi)$ then either $\mathbf{EA} \vdash K(\varphi)$ or $\mathbf{EA} \vdash K(\psi)$.

c. Modal epistemic arithmetic (MEA)

In Shapiro's $\mathbf{EA}$, K is a primitive operator which cannot be further analyzed. Horsten (1994) suggests that the provability operator is not primitive but complex. He distinguishes between two components of informal provability: the modal and the epistemic.

The modal component is associated with possibility. The epistemic component is explained in terms of a mathematical proof. Instead of just one operator K we have two unary operators applying to formulas: $\Diamond$ and P, where $\Diamond$ is interpreted as possibility and P intuitively stands for "some mathematician has a proof that...". In $\mathcal{L}_{\mathbf{PA}}$ extended with these two operators, $\mathcal{L}_{MEA}$, and following these intuitions we present the so-called Modal Epistemic Arithmetic (**MEA**) (Horsten, 1994). The axioms of **MEA** are as follows:

1. all the axioms of **PA** with induction for the extended language,
2. $\Diamond\varphi \rightarrow \varphi$ where φ is ontic i.e. $\varphi \in \mathcal{L}_{\mathbf{PA}}$,
3. $P(\varphi) \rightarrow \varphi$,
4. $P(\varphi) \rightarrow P(P(\varphi))$,
5. $(\Diamond P(\varphi) \wedge \Diamond P(\varphi \rightarrow \psi)) \rightarrow \Diamond P(\psi)$,
6. all axioms of the modal system **S5** for $\Diamond$,

and a rule of inference: if φ is a theorem, then so is $\Diamond P(\varphi)$.

Axioms 1 and 2 are some variants of the reflection principle which is provable for P for ontic sentences, and for $\Diamond$ for all sentences. It does not follow that reflection is provable for $\Diamond P$. Axioms 3 and 4 are standard axioms for provability ((HB3) and (HB1)). Note that (HB3) works for provability operator and (HB1) for $\Diamond P$. By a $\Diamond P$-formula we will mean any formula φ where all subformulas of φ of the form $P\chi$ are immediately preceded with $\Diamond$.

Observation 5.5. *Let $\varphi \in \mathcal{L}_{MEA}$ be a $\Diamond P$-formula. Then the following claims hold:*

$$\mathbf{MEA} \vdash \Diamond P\varphi \to \varphi$$
$$\mathbf{MEA} \vdash \Diamond P\varphi \to \Diamond P \Diamond P\varphi$$

The above observation shows that we have a certain version of reflection schema and certain version of (HB3), at least for a restricted class of formulas.

d. Problems with provability as an operator

The main aim of treating provability as an operator is to circumvent the impossibility that arises for the formal provability predicate — that of having all HB conditions and all the instances of the reflection schema at the same time.

Theorem 5.6 (Montague's theorem)**.** *Peano Arithmetic, if consistent, cannot contain (or be consistently extended to contain) a (possibly complex) predicate for which all Hilbert-Bernays conditions and all instances of the reflection schema hold.*

Proof. Suppose that there is such a predicate, call it P. We will use natural deduction system. Argue inside the theory:

1. $\lambda \equiv P(\ulcorner \neg\lambda \urcorner)$	Diagonal lemma
1.1 λ	Hypothesis
1.2 $P(\ulcorner \neg\lambda \urcorner)$	equivalence elimination: 1,1.1
1.3 $\neg\lambda$	modus ponens and reflection schema: 1.2
2. $\neg\lambda$	reductio ad absurdum: 1.1 → 1.3
3. $P(\ulcorner \neg\lambda \urcorner)$	(HB1)
4. $\neg P(\ulcorner \neg\lambda \urcorner)$	CL, 1, 2
5. contradiction	CL, 3, 4

□

In order to prove Montague's theorem one applies the diagonal lemma to a certain formula involving provability predicate. But if provability is treated as an operator, we cannot use the diagonal lemma to generate this pardoxical formula.

MEA is capable of proving variants of reflection schema. It is an interesting result, for the name of the game here is to gather as many instances of reflection schema as possible without inconsistency. Unfortunately, the theory has some other philosophical problems:

1. The choice which rules are postulated for P and which are postulated for $\Diamond$ seems somewhat arbitrary. It is possible to consider different combination of those rules. For instance, to add axiom (K) directly for P.

2. The reflection schema is available only for $\Diamond P$. It is not clear why other types of reflection shouldn't be introduced. For instance, reflection restricted to Σ_1 formulas doesn't look completely insane.

3. Usually provability is treated as a predicate and not as an operator. There seems to be no motivation for using an operator, independent of blocking the Montague's theorem.

4. Both **EA** and **MEA** seem to be a bit too weak– there are translations to **HA** which preserve theorems.

e. PEA and its basis

Another strategy is to treat informal provability as a predicate and weaken some of the Hilbert-Bernays conditions for this predicate. Again, expand $\mathcal{L}_{\mathbf{PA}}$ with an additional predicate P for informal provability, thus obtaining a new language $\mathcal{L}_P$. The idea here is straightforward: we divide the set of problematic principles (HB conditions and the reflection schema) for the additional predicate P between two theories: PPEA and its basis. Then we add to PEA all the instances of the axiom saying that if something is derivable in the basis, it is informally provable.

We will start with a theory called the basis of **PEA** (**BPEA**) (Horsten, 1997), which is defined by:

Basis Axiom 1 $\mathbf{PA}$ in extended language with induction extended to $\mathcal{L}_P$
Basis Axiom 2 $P(\ulcorner\varphi\urcorner) \to (P(\ulcorner\varphi \to \psi\urcorner) \to P(\ulcorner\psi\urcorner))$ for all $\varphi, \psi \in \mathcal{L}_P$
Basis Axiom 3 $P(\ulcorner\varphi\urcorner) \to P(\ulcorner P(\ulcorner\varphi\urcorner)\urcorner)$ for all $\varphi \in \mathcal{L}_P$

So, we have (K) and (4) for P. By $\mathtt{Prov}_B$ we mean the standard provability predicate of **BPEA**. **PEA** is given by the following axioms:

Axiom 1 $\mathbf{PA}$ in the extended language with induction extended to $\mathcal{L}_P$
Axiom 2 $P(\ulcorner\varphi\urcorner) \to \varphi$ for all $\varphi \in \mathcal{L}_P$
Axiom 3 $\mathtt{Prov}_B(\ulcorner\varphi\urcorner) \to P(\ulcorner\varphi\urcorner)$ for all $\varphi \in \mathcal{L}_P$

We have the reflection schema for P. Notice that we do not have (Nec) for P, but we have the implication $\mathtt{Prov}_B(\ulcorner\varphi\urcorner) \to P(\ulcorner\varphi\urcorner)$, which together with the reflection schema gives us $\mathtt{Prov}_B(\ulcorner\varphi\urcorner) \to \varphi$ which is a certain version of (Nec).

These theories are still under investigation. One of the nice things about **PEA**, apart from the reflection schema holding in it, is the fact that **PEA** has nice models.

Fact 5.7. *PEA has a model based on the standard model of arithmetic.*

However, it seems that the philosophical motivations underlying the system are somewhat lacking. While informal provability seems unified, this system clearly has two separate layers. The restrictions on the claims for which reflection can be used is still there — it's just that they're somewhat less visible, because they arise at the point in which a restriction is put on what can be provably provable (**Axiom 3**). Yes, **Axiom 2** guarantees that reflection is provable for any ϕ, but given that the internal logic of P is built starting from the forma provability predicate of **BPEA**, it holds universally at the price of being idle on many occasions.

f. Non-deterministic many-valued approach

Another, rather non-standard approach (Pawlowski and Urbaniak, 2017) is to change the underlying logic and to build theories of informal provability where the notion is treated as a partial notion. The partition seems intuitive: some mathematical sentences are informally provable, others are informally refutable and some, it seems, are informally undecidable.

In order to model reasoning with a partial notion formally a three-valued logic comes handy. So think about partitions as values: 1 stands for *informally provable*, 0 for *informally refutable*, and n for *neither*.

One initial problem is that if we take a close look at disjunctions or conjunctions of mathematical sentences, it seems that their logical value depends not only on the values of their disjuncts or conjuncts. For instance, take two disjunction of two undecidable sentences. One is built from the Continuum Hypothesis and its negation. The other one is composed from the claim that the standard set theory (ZFC) is consistent, and the General Continuum hypothesis. The first one is informally provable, because it is a substitution of propositional tautology, whereas the other disjunction at least isn't known to be informally provable, and there is no contradiction in thinking that it is not informally provable.

In order to formally represent the above intuition consider a standard propositional language with one additional modal operator B ($\mathcal{L}_B$), intuitively read as *it is provable that*. Use the three values and the indeterminacy discussed above to define matrices for connectives and the operator B by:

φ	ψ	$\neg\varphi$	$B\varphi$	$\varphi \vee \psi$	$\varphi \wedge \psi$	$\varphi \to \psi$	$\varphi \equiv \psi$
1	1	0	1	1	1	1	1
n	1	n	$0/n$	1	n	1	n
0	1	1	0	1	0	1	0
1	n			1	n	n	n
n	n			$n/1$	$0/n$	$n/1$	$0/n/1$
0	n			n	0	1	n
1	0			1	0	0	0
n	0			n	0	n	n
0	0			0	0	1	1

where for two values x, y, when we write x/y we mean that either the formula has value x or y. The matrix is rather straightforward. The only interesting case is for B when φ has value n. Then either we cannot informally prove its undecidability then it remains n or we can do that, at the same time disproving $B(\varphi)$, hence value 0.

By an *assignment* we mean a function $v : Prop \mapsto Val$ from the set of propositional variables to the set of values. An *evaluation* is an extension of the assignment to complex formulas respecting conditions given above. The general phenomenon is that an assignment doesn't unambiguously determine unique evaluation: it only underlies a class of evaluations that extend it. For instance, if $v(p) = v(q) = n$, there will be one evaluation with $e^1_v(p \vee q) = n$ and another with $e^2_v(p \vee q = 1)$.

If we were to define a logic in terms of preservation of value 1, it would turn out to be too weak (for instance, conjunction and disjunction aren't even guaranteed to commute). We need one more requirement:

Definition 5.8 (Closure condition)**.** For any $\mathcal{L}_{\mathcal{B}}$-formulas $\varphi_1, \varphi_2, \cdots, \varphi_n, \psi$ such that

$$\varphi_1, \varphi_2, \cdots, \varphi_n \models \psi,$$

where $\models$ is the classical consequence relation for $\mathcal{L}_B$, for any e, if $e(B\varphi_i) = 1$ for any $0 < i \leq n$, then $e(B\psi) = 1$.

We will use 🦇$_C$ φ iff for all evaluations satisfying the closure condition φ has value 1. Similarly, we define Γ🦇$_C$ φ as the preservation of value 1 in all evaluations satisfying the closure condition. This logic is called **CABAT**.

One of the most interesting features of this logic is the fact that under a certain translation of the standard provability predicate as B, the translation of Löb's theorem doesn't hold. This is a good sign. There was no initial intuition that Löb's theorem is correct principle for informal provability. It's a rather unwanted technical result.

If we look at Montague's theorem, after applying the diagonal lemma, the rest of the proof is done on the propositional level. We can translate all the premises of the theorem together with the formula resulting from application of diagonal lemma into CABAT langauge. It turned out that the theorem does not hold. This shows that using quite natural philosophical intuitions we can build a formal system with certain intuitive principles for provability which can be consistently extended with all the instances of the reflection schema.

The above proposal has its drawbacks. It is nothing but obscure how to build an arithmetical theory using this logic. It seems that the most common strategies for building partial models using three-valued logics will not work here. The second issue is that this proposal is still underdeveloped. The current semantics for CABAT is convoluted – it is not clear which evaluations are removed by the closure condition. Simpler and more intuitive semantics needs to be developed. Moreover, provability here is treated as an operator not as a predicate, and it is not clear what the consequences of moving to the predicate level with this logic would be.

g. Conditional epistemic obligation and believability logic

A somewhat different formal approach to our (at least prima facie) commitment to reflection arose in the context of formal axiomatic theories of truth built over arithmetic (see Horsten, 2011; Halbach, 2011; Cieśliński, 2016a, for more details on the truth-theoretic aspecs of the developments).

Truth-theoretic considerations aside, from the axioms and rules of $\mathbf{PA}$ local reflection for $\mathbf{PA}$ doesn't follow, and so we don't seem logically commited to local reflection for $\mathbf{PA}$ just because we accept the axioms and rules of the system. Yet, there seems something irrational about accepting $\mathbf{PA}$ without thinking that for any $\varphi \in \mathcal{L}_{\mathbf{PA}}$, indeed, if $\mathbf{PA} \vdash \varphi$, then φ (if we had a truth predicate available, we could use a single claim: that *whatever* $\mathbf{PA}$ *proves, is true*; but we're trying to avoid getting into a discussion of theories of truth). Assuming this is correct, the challenge is to explain why someone who accepts $\mathbf{PA}$ is rationally commited to reflection which nevertheless doesn't logically follow from the axioms of $\mathbf{PA}$ by the rules of $\mathbf{PA}$. (Another example of a commitment of this sort is that to the Gödel sentence of $\mathbf{PA}$, which seems true, even though it doesn't follow from $\mathbf{PA}$.)

For such occasions, Ketland (2005) introduced the notion of a *conditional epistemic obligation*. Ketland hasn't really explicated the notion, but only pointed out that once we accept a theory, we become conditionally epistemically obligated to accept some other claims in its language which nevertheless don't follow from the theory itself, and listed some examples such as that of reflection or the Gödel sentence.

A philosophically interesting explication of the notion of conditional epistemic obligation is not trivial (see Cieśliński, 2016a,b, for a discussion). But even putting this daunting task aside, the question arises whether we can achieve a more humble goal: that of describing the inferential behavior of the predicate expressing this sort of commitment by means of a formal system. Such an attempt can be found in the works of Cieślinski. For the sake of simplicity, we'll discuss the system as built over a particular arithmetical theory, **PA**.

Extend $\mathcal{L}_{\mathbf{PA}}$ with a new unary predicate symbol B, thus obtaining a new language $\mathcal{L}_B$. The goal is to describe the *theory of believabity* built over **PA**. Let the result of taking the axioms of $\mathbf{PA}$ with induction extended to $\mathcal{L}_B$ be called **PAB**. Theory **Bel(PA)** extends **PAB** with the following axioms:

$$\begin{array}{ll}(\mathrm{A}_1) & \forall\psi \in \mathcal{L}_B\ [\mathtt{Prov}_{PAB}(\ulcorner\psi\urcorner) \to B(\ulcorner\psi\urcorner)] \\ (\mathrm{A}_2) & \forall\varphi,\psi \in \mathcal{L}_B\ [B(\ulcorner\varphi\urcorner) \wedge B(\ulcorner\varphi \to \psi\urcorner) \to B(\ulcorner\psi\urcorner)]\end{array}$$

On top of the axioms, **Bel(PA)** has two additional rules of inference. (Nec) for B, which allows to infer

$$\mathbf{Bel}(\mathbf{PA}) \vdash B(\ulcorner\psi\urcorner)$$

from

$$\mathbf{Bel}(\mathbf{PA}) \vdash \psi,$$

and the *generalization rule* (Gen), which allows to infer

$$\mathbf{Bel}(\mathbf{PA}) \vdash B(\ulcorner\forall x\ \psi(x)\urcorner)$$

from

$$\mathbf{Bel}(\mathbf{PA}) \vdash \forall x\ B(\ulcorner\psi(x)\urcorner).$$

While the motivations for (Nec) and the axioms are quite straightforward (and not completely different from the considerations pertaining to **PEA**), one thing that makes the theory stand apart is (Gen). Normally, in **PA**, just because for every standard numeral $\overline{n}$ $\mathbf{PA} \vdash \psi(\overline{n})$, it doesn't follow that $\mathbf{PA} \vdash \forall x\ (\psi(x))$. This is because $\mathbf{PA}$ as a first-order theory admits non-standard models, and so can have a model in which there are non-standard numbers not denoted by any standard numeral. It is exactly (Gen) that allows our commitment (tracked by B, that is, the internal logic of **Bel(PA)**) to go beyond what $\mathbf{PA}$ already can prove, including reflection and the consistency of **PA**.

6 Further topics

The scope of this survey (which is, admittedly, already quite long) is limited. In this section we list and briefly describe multiple further issues related to provability, which we couldn't properly cover. The list is, of course, far from complete.

a. Mixed logic of consistency and ω-consistency

One interesting fact about the interaction between the notions of consistency and ω consistency is that the negation of the consistency of **PA**, while not being inconsistent with **PA**, is ω-inconsistent with **PA**. Now the question is: can we develop a propositional logic to reason about such claims?

For this purpose we need a *bimodal* logic with two modalities (Ignatiev, 1993; Japaridze, 1985). Let $\Box\psi$ stand for the provability of ψ, and $\blacksquare$ stand for the ω-inconsistency of $\neg\psi$ with the background theory.

The axioms of **GLB** (**B** from *bimodal*) are all tautologies, all instances of (K) for $\Box$, all instances of (K) for $\blacksquare$, all instances of (Löb) for $\Box$ and all instances of (Löb) for $\blacksquare$, and all instances of the following schemata:

$$\Box\varphi \to \blacksquare\varphi$$
$$\neg\Box\varphi \to \blacksquare\neg\Box\varphi$$

The rules are *modus ponens* and (Nec) for $\Box$ (clearly, (Nec) for $\blacksquare$ follows by the first of the above formulas). Given a realization r for the language of **GLB**, it is extented to a **T**-interpretation by the standard conditions for classical connectives together with:

$$r_{\mathbf{T}}(\Box\varphi) = \mathtt{Prov}_{\mathbf{T}}(\ulcorner r_{\mathbf{T}}(\varphi)\urcorner)$$
$$r_{\mathbf{T}}(\blacksquare\varphi) = \neg\omega\mathtt{Con}_{T}(\ulcorner r_{\mathbf{T}}(\neg\varphi)\urcorner)$$

GLB is arithmetically sound and complete.

b. Provability logic with quantifiers

One way to extend $\mathcal{L}_M$ with quantifiers is to add *propositional quantifiers* binding propositional variables, which would allow to express claims such as 'some formula is not provable', or 'for every formula there is one which is provable just in case the former one isn't' (by $\exists p\ \neg\Box p$ and $\forall p\ \exists q\ (\Box q \equiv \neg\Box p)$. As proven by (Shavrukov, 1997), the set of arithmetically valid sentences of this language is undecidable.

Another move to consider is moving to a first-order modal language and extending the intended semantics appropriately. Alas, the set of first-order formulas true in every realization is not effectively axiomatizable (Artemov, 1985), and neither is the set of formulas provable in **PA** under any realization (the complexity of this set Π^0_2) (Vardanyan, 1986). Montagna et al. (1984) showed moreover that quantified **GL** is not complete with respect to any class of Kripke frames, and that it doesn't have the fixed point property. Similarly, the first-order version of the logic of proofs has been proven to not be recursively enumerable (Yavorsky, 2001).

c. Interpretability logics

The notion of interpretability was introduced into meta-logic by Tarski et al. (1953). A theory **T** is interpretable in theory **U** just in case the language of **T** can be translated into that of **U** so that the translations of theorems of **T** become theorems of **U**. One example of interpretability is the relation between **PA** and the standard set theory **ZFC**.

There is a translation from the language of arithmetic into the language of set theory, such that the translations of all theorems of $\mathbf{PA}$ are provable in **ZFC**.

The formal symbol representing interpretability was introduced into a language of a logic of provability by Švejdar (1983). The intended reading of $\varphi \rhd \psi$ if that for a sufficiently arithmetically rich theory $\mathbf{T}$ (such as **[PA]**), $\mathbf{T}+\psi$ is interpretable in $\mathbf{T}+\varphi$.

Interpretability logics were further studied by Visser (1990, 1998). There is a sensible logic $\mathbf{IL}$ which axiomatizes interpretability principles valid in all sensible theories. It is $\mathbf{GL}$ expanded with:

$$\begin{aligned}
\Box(\varphi \to \psi) &\to \varphi \rhd \psi \\
(\varphi \rhd \psi \wedge \psi \rhd \chi) &\to \varphi \rhd \chi \\
(\varphi \rhd \chi \wedge \psi \rhd \chi) &\to (\varphi \vee \psi) \rhd \chi \\
\varphi \rhd \psi &\to (\Diamond\varphi \to \Diamond\psi) \\
(\Diamond\varphi) &\rhd \varphi
\end{aligned}$$

However, the class of all principles that hold in all realizations is sensitive to the choice of the underlying theory (see Visser, 1998, for a comprehensive survey).

d. Generalization and classifications

The notion of provability can be considered on a more general level, that of *provability logic of a given theory* $\mathbf{T}$ *relative to a metatheory* $\mathbf{U}$ - the notion was introduced by Artemov (1987) and Visser (1981). Such a logic, denoted by $\mathbf{PL_T}(\mathbf{U})$, is the set of all propositional principles of provability for $\mathbf{T}$ that can be proven in $\mathbf{U}$. From this perspective, $\mathbf{GL}$ is the provability logic $\mathbf{PL_{PA}}(\mathbf{PA})$, and **S** is $\mathbf{PL_{PA}}(Tr(\mathbb{N}))$, where $Tr(\mathbb{N})$ is the set of all $\mathcal{L}_{\mathbf{PA}}$-formulas true in the standard model of arithmetic. Much work has been done on the classification of logics $\mathbf{PL_T}(\mathbf{U})$, where $\mathbf{T}$ and $\mathbf{U}$ are known and independently studied extensions of $\mathbf{PA}$ (Artemov, 1987, 1986; Beklemishev, 1990; Japaridze, 1986; Visser, 1984).

e. Algebraic approaches

Magari (1975a,b) developed an algebraic approach to provability logic. The Magari algebra of **T**, is the set of **T**-sentences factorized modulo equivalence within **T**. Further applications of the algebraic toolkit to provability logics are (Montagna, 1978, 1979).

f. Connections with other domains

Provability logics have some use in computability theory. For instance, Beklemishev (1990) uses them to investigate which computable functions can be proved to be total by means of restricted induction schemata. Another domain where provability logics find applications is proof theory (Beklemishev, 2003); see also (Japaridze and Jongh, 1998) for a survey.

7 References and further readings

A historical account of the beginnings of the development of the logics of provability can be found in (Boolos and Sambin, 1991). For more introductory surveys of the logics of provability, read (Verbrugge, 2016), (Švejdar, 2000). For a short, but dense survey see (Artemov, 2006). For a survey focusing on self-reference read (Smoryński, 2004). For material focusing on introducing the Logic of Proofs consult (Artemov, 2007). For more advanced surveys of the logic of provability, consult (Japaridze and Jongh, 1998) and (Artemov and Beklemishev, 2005). As for full blown book-long treatments, (Boolos, 1993) is invaluable, and so is (Smoryński, 1985).

References

Anderson, A. R. (1956). The formal analysis of normative systems. Technical report, DTIC Document.

Artemov, S. (1985). Nonarithmeticity of truth predicate logics of provability. *Doklady Akademii Nauk SSSR*, 284(2):270–271.

Artemov, S. (1986). On modal logics axiomatizing provability. *Mathematics of the USSR-Izvestiya*, 27(3):401–429.

Artemov, S. (1987). Arithmetically complete modal theories. *American Mathematical Society Translations*, 2(135):39–54.

Artemov, S. (1994). Logic of proofs. *Annals of Pure and Applied Logic*, 67(1-3):29–59.

Artemov, S. (1995). Operational modal logic. Technical report, Cornell University, MSI 95–29.

Artemov, S. (1998). Logic of proofs: a unified semantics for modality and λ-terms. Technical report, Cornell University, CFIS 98–06.

Artemov, S. (2001). Explicit provability and constructive semantics. *Bulletin of Symbolic logic*, pages 1–36.

Artemov, S. (2006). Modal logic in mathematics. In Blackburn, P., van Benthem, J., and Wolter, F., editors, *Handbook of Modal Logic*, pages 927–970. Elsevier.

Artemov, S. (2007). On two models of provability. In *Mathematical problems from applied logic II*, pages 1–52. Springer.

Artemov, S. N. and Beklemishev, L. D. (2005). Provability logic. In *Handbook of Philosophical Logic, 2nd Edition*, pages 189–360. Springer.

Beklemishev, L. D. (1990). On the classification of propositional provability logics. *Mathematics of the USSR-Izvestiya*, 35(2):247–275.

Beklemishev, L. D. (2003). Proof-theoretic analysis by iterated reflection. *Archive for Mathematical Logic*, 42(6):515–552.

Blackburn, P., de Rijke, M., and Venema, Y. (2001). *Modal Logic*. Cambridge University Press.

Boolos, G. (1993). *The Logic of Provability*. Cambridge University Press.

Boolos, G. and Sambin, G. (1991). Provability: the emergence of a mathematical modality. *Studia Logica*, 50(1):1–23.

Carnap, R. (1934). *Logische Syntax der Sprache*. Springer Berlin Heidelberg, Berlin, Heidelberg.

Cieśliński, C. (2016a). *The Epistemic Lightness of Truth. Deflationism and its Logic*. Cambridge University Press, [forthcoming].

Cieśliński, C. (2016b). Minimalism and the generalisation problem: on Horwich's second solution. *Synthese*, pages 1–25. DOI 10.1007/s11229-016-1227-5.

Feferman, S., Dawson Jr, J. W., Kleene, S. C., Moore, G. H., Solovay, R. M., and van Heijenoort, J., editors (1986). *Kurt Gödel: collected works. Vol. 1: Publications 1929-1936*. Oxford University Press.

Fitting, M. (2003). A semantics for the logic of proofs. Technical report, CUNY Ph. D. Program in Computer Science, TR - 2003012.

Flagg, R. C. and Friedman, H. (1986). Epistemic and intuitionistic formal systems. *Annals of Pure and Applied Logic*, 32(1):53–60.

Gaifman, H. (2006). Naming and diagonalization, from cantor to gödel to kleene. *Logic Journal of the IGPL*, 14(5):709–728.

Gödel, K. (1933). Eine Interpretation des intuitionistischen Aussagenkalkuls. [reprinted in (Feferman et al., 1986)].

Goodman, N. D. (1984). Epistemic arithmetic is a conservative extension of intuitionistic arithmetic. *Journal of Symbolic Logic*, 49(1):192–203.

Grzegorczyk, A. (1967). Some relational systems and the associated topological spaces. *Fundamenta Mathematicae*, 60(3):223–231.

Hájek, P. and Pudlak, P. (1993). Metamathematics of first-order arithmetic. *Berlin: Springer*, 2:295–297.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge University Press.

Hedman, S. (2004). *A first course in logic : an introduction to model theory, proof theory, computability, and complexity*. Oxford University Press, Oxford New York.

Henkin, L. (1952). A problem concerning provability. *Journal of Symbolic Logic*, 17(2):160.

Heyting, A. (1930). *Die formalen Regeln der intuitionistischen Mathematik*. Verlag der Akademie der Wissenschaften.

Heyting, A. (1931). Die intuitionistische Grundlegung der Mathematik. *Erkenntnis*, 2(1):106–115.

Heyting, A. (1934). *Mathematische Grundlagenforschung Intuitionismus, Beweistheorie*. Springer.

Hilbert, D. and Bernays, P. (1939). *Grundlagen der Mathematik II*. Springer.

Horsten, L. (1994). Modal-epistemic variants of Shapiro's system of epistemic arithmetic. *Notre Dame Journal of Formal Logic*, 35(2):284–291.

Horsten, L. (1997). Provability in principle and controversial constructivistic principles. *Journal of Philosophical Logic*, 26(6):635–660.

Horsten, L. (1998). In defence of epistemic arithmetic. *Synthese*, 116:1–25.

Horsten, L. (2011). *The Tarskian Turn. Deflationism and Axiomatic Truth*. MIT Press.

Ignatiev, K. N. (1993). On strong provability predicates and the associated modal logics. *The Journal of Symbolic Logic*, 58(01):249–290.

Japaridze, G. and Jongh, D. d. (1998). The logic of provability. In Buss, S. R., editor, *Handbook of proof theory*, volume 137, pages 475–550. Elsevier.

Japaridze, G. K. (1985). The polymodal logic of provability. In *Intensional Logics and Logical Structure of Theories: Material from the fourth Soviet–Finnish Symposium on Logic, Telavi*, pages 16–48.

Japaridze, G. K. (1986). *The modal logical means of investigation of provability*. PhD thesis, Moscow State University.

Kaye, R. (1991). *Models of Peano arithmetic*. Oxford University Press.

Ketland, J. (2005). Deflationism and the gödel phenomena: reply to tennant. *Mind*, 114(453):75–88.

Kolmogorov, A. (1932). Zur Deutung der Intuitionistischen Logik. *Mathematische Zeitschrift*, 35(1):58–65.

Kossak, R. (2006). *The structure of models of Peano arithmetic*. Clarendon, Oxford.

Leitgeb, H. (2009). On formal and informal provability. In *New Waves in Philosophy of Mathematics*, pages 263–299. New York: Palgrave Macmillan.

Löb, M. H. (1955). Solution of a problem of Leon Henkin. *The Journal of Symbolic Logic*, 20(02):115–118.

Magari, R. (1975a). The diagonalizable algebras (the algebraization of the theories which express Theor.:ii). *Bollettino della Unione Matematica Italiana*, 4(12):117–125.

Magari, R. (1975b). Representation and duality theory for diagonalizable algebras. *Studia Logica*, 34(4):305–313.

Marfori, M. A. (2010). Informal proofs and mathematical rigour. *Studia Logica*, 96:261–272.

McKinsey, J. C. and Tarski, A. (1948). Some theorems about the sentential calculi of Lewis and Heyting. *The Journal of Symbolic Logic*, 13(01):1–15.

Mkrtychev, A. (1997). Models for the logic of proofs. In Adian, S. and Nerode, A., editors, *Logical Foundations of Computer Science '97*, pages 266–275. Springer.

Montagna, F. (1978). On the algebraization of a feferman's predicate. *Studia Logica*, 37(3):221–236.

Montagna, F. (1979). On the diagonalizable algebra of peano arithmetic. *Bollettino della Unione Matematica Italiana*, 16(5):795–812.

Montagna, F. et al. (1984). The predicate modal logic of provability. *Notre Dame Journal of Formal Logic*, 25(2):179–189.

Myhill, J. (1960). Some remarks on the notion of proof. *Journal of Philosophy*, 57(14):461–471.

Nogina, E. (1994). Logic of proofs with the strong provability operator. Technical report, Institute for Logic, Language and Computation, University of Amsterdam, ILLC Prepublication Series ML-94-10.

Nogina, E. (1996). Grzegorczyk logic with arithmetical proof operators. *Fundamentalnaya i Prikladnaya Matematika*, 2(2):483–499.

Nowell-Smith, P. and Lemmon, E. (1960). Escapism: the logical basis of ethics. *Mind*, 69(275):289–300.

Orlov, I. E. (1928). The calculus of compatibility of propositions. *Mathematics of the USSR, Sbornik*, 35:263–286.

Pawlowski, P. and Urbaniak, R. (2017). Many-valued logic of informal provability: a non-deterministic strategy. *The Review of Symbolic Logic*. [accepted and forthcoming].

Segerberg, K. K. (1971). *An essay in classical modal logic*. The Philosophical Society in Uppsala.

Shapiro, S. (1985). Epistemic and intuitionistic arithemtic. In *Intensional mathematics*. North Holland.

Shavrukov, V. Y. (1997). Undecidability in diagonalizable algebras. *The Journal of Symbolic Logic*, 62(01):79–116.

Simpson, S. G. (2009). *Subsystems of second order arithmetic*, volume 1. Cambridge University Press.

Smiley, T. J. (1963). The logical basis of ethics. *Acta Philosophica Fennica*, 16:237–246.

Smith, P. (2007). *An Introduction to Gödel's Theorems*. Cambridge University Press.

Smoryński, C. (1985). *Self-reference and modal logic*. Universitext. Springer, New York.

Smoryński, C. (2004). Modal Logic and Self-Reference. In Gabbay, D. and Guenthner, F., editors, *Handbook of Philosophical Logic*, volume 11, pages 1–53. Springer.

Solovay, R. M. (1976). Provability interpretations of modal logic. *Israel Journal of Mathematics*, 25(3-4):287–304.

Sundholm, G. (1998). Proofs as acts and proofs as objects: some questions for Dag Prawitz. *Theoria*, 64(2-3):187–216.

Švejdar, V. (1983). Modal analysis of generalized Rosser sentences. *The Journal of Symbolic Logic*, 48(04):986–999.

Švejdar, V. (2000). On provability logic. *Nordic Journal of Philosophical Logic*, 4(2):95–116.

Tarski, A., Mostowski, A., and Robinson, R. M. (1953). *Undecidable theories*. Elsevier.

Troelstra, A. and van Dalen, D. (1988). *Constructivism in Mathematics*, volume 1 and 2. Elsevier.

Vardanyan, V. A. (1986). Arithmetic complexity of predicate logics of provability and their fragments. *Soviet Mathematics Doklady*, 33(3):569–572.

Verbrugge, R. L. (2016). Provability logic. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Spring 2016 edition.

Visser, A. (1981). *Aspects of Diagonalization and Probability*. PhD thesis, University of Utrecht.

Visser, A. (1984). The provability logics of recursively enumerable theories extending Peano Arithmetic at arbitrary theories extending Peano Arithmetic. *Journal of Philosophical Logic*, 13(1):97–113.

Visser, A. (1990). Interpretability logic. In Petkov, P., editor, *Mathematical logic*, pages 175–209. Springer.

Visser, A. (1997). A course on bimodal provability logic. annals of pure and applied logic, vol. 73 (1995), pp. 109–142. *The Journal of Symbolic Logic*, 62(02):686–687.

Visser, A. (1998). An overview of interpretability logic. In Kracht, M., de Rijke, M., Wansing, H., and Zakharyaschev, M., editors, *Advances in Modal Logic*, volume 1. CSLI Publications.

Visser, A. (2008). Propositional combinations of σ-sentences in heyting's arithmetic. *Logic Group Preprint Series*, 117.

Yavorskaya, T. (2001). Logic of proofs and provability. *Annals of pure and applied logic*, 113(1):345–372.

Yavorsky, R. E. (2001). Provability logics with quantifiers on proofs. *Annals of Pure and Applied Logic*, 113(1):373–387.