The Interpretability Logic of Peano Arithmetic

Alessandro Berarducci

The Journal of Symbolic Logic, Vol. 55, No. 3. (Sep., 1990), pp. 1059-1089.

Stable URL:
http://links.jstor.org/sici?sici=0022-4812%28199009%2955%3A3%3C1059%3ATILOPA%3E2.0.CO%3B2-J

The Journal of Symbolic Logic is currently published by Association for Symbolic Logic.

THE JOURNAL OF SYMBOLIC LOGIC
Volume 55, Number 3, Sept. 1990

THE INTERPRETABILITY LOGIC OF PEANO ARITHMETIC

ALESSANDRO BERARDUCCI

To Silvia

Abstract. PA is Peano arithmetic. The formula $\mathrm{Interp}_{\mathrm{PA}}(\alpha, \beta)$ is a formalization of the assertion that the theory PA + α interprets the theory PA + β (the variables α and β are intended to range over codes of sentences of PA). We extend Solovay's modal analysis of the formalized provability predicate of PA, $\mathrm{Pr}_{\mathrm{PA}}(x)$, to the case of the formalized interpretability relation $\mathrm{Interp}_{\mathrm{PA}}(x, y)$. The relevant modal logic, in addition to the usual provability operator '$\Box$', has a binary operator '$\triangleright$' to be interpreted as the formalized interpretability relation. We give an axiomatization and a decision procedure for the class of those modal formulas that express valid interpretability principles (for every assignment of the atomic modal formulas to sentences of PA). Our results continue to hold if we replace the base theory PA with Zermelo-Fraenkel set theory, but not with Gödel-Bernays set theory. This sensitivity to the base theory shows that the language is quite expressive. Our proof uses in an essential way earlier work done by A. Visser, D. de Jongh, and F. Veltman on this problem.

§1. Introduction. The relation of relative interpretability between axiomatic theories (formulated in first order logic) has been used to prove relative consistency results, decidability and undecidability of theories, and to compare the strength of theories.

Intuitively 'T interprets S' means that the language of S is translatable into the language of T in such a way that T proves the translation of every axiom of S (as in §2.1). For example Zermelo-Fraenkel set theory (ZF) interprets Peano arithmetic (PA).

Let T be a theory which has a reasonable notion of natural numbers and finite sequences (say T is what Pudlák [11, p. 435] calls a 'sequential theory'). Suppose also that T satisfies 'full induction', namely, for every formula $\theta(x)$ in the language of T, T proves that if there is a natural number satisfying the formula $\theta(x)$, then there is a least such number. Typical examples are ZF and PA, but not Gödel-Bernays set theory GB (which does not satisfy full induction). By (an easy generalization of) a theorem of Orey, T interprets S if and only if for every finite subtheory S' of S, T proves that S' is consistent (here S is a recursively axiomatized theory). It follows

Received April 21, 1989; revised September 1, 1989.
Partially supported by NSF grant DMS 8701828.

0022-4812/90/5503-0012/$04.10

in particular that if a sequential theory T satisfies full induction, then T is not finitely axiomatizable. (Proof: clearly T interprets T. If T were finitely axiomatizable, then by Orey's theorem T would prove the consistency of itself, contradicting Gödel's second incompleteness theorem.)

For finitely axiomatizable sequential theories there is a theorem of Harvey Friedman (see [7] or [14]) that gives a similar characterization of interpretability in terms of consistency: let T and S be finitely axiomatized sequential theories, then T interprets S iff the theory $\mathrm{I}\Delta_0 + \mathrm{EXP}$ proves that the consistency of T (with respect to cut-free proofs) implies the consistency of S (with respect to cut-free proofs). We recall that '$\mathrm{I}\Delta_0$' is the fragment of PA obtained by restricting the induction scheme to Δ_0-formulas, and 'EXP' is the assertion that the exponential function '2^x' is total.

Friedman's and Orey's theorems provide a characterization of interpretability for a large class of sequential theories, but they do not say anything about theories which do not have a good notion of finite sequences and natural numbers, for example the theory of real closed fields (which is not in the scope of our investigation).

In this paper we use the language of modal logic to give an axiomatic treatment of interpretability similar to and extending the corresponding axiomatic treatment of provability carried out by Solovay in [16].

We need to fix a 'base' theory T satisfying the hypothesis of Orey's theorem and such that T does not prove false Σ^0_1-assertions (e.g. PA or ZF). Any such theory will work, but for simplicity we take PA as our base theory.

The modal language of interpretability, in addition to the usual modal operator '$\Box$', has a binary modal operator '$\rhd$' standing for the relation of interpretability over PA. More precisely we consider interpretations of the modal language into the language of PA such that if the modal formulas A and B are interpreted as the PA-sentences α and β, then the formula $A \rhd B$ is interpreted as a formalization of the assertion '$\mathrm{PA} + \alpha$ interprets $\mathrm{PA} + \beta$', and the modal formula $\Box A$ is interpreted as a formalization of 'PA proves α'.

We make use of Orey's theorem to give an axiomatization and a decision procedure for the class of those modal formulas that express valid interpretability and provability principles (for every assignment of the atomic modal formulas to sentences of PA).

It turns out that if one replaces the base theory PA with a (strong enough) finitely axiomatizable sequential theory (e.g. ACA_0 of GB), then one gets a different modal logic. (The completeness of this logic with respect to interpretations in ACA_0 and GB has been proved by A. Visser and C. Smoryński and uses Friedman's characterization of interpretability.)

For example the modal formula $A \rhd B \to \Box(A \rhd B)$ is valid if the base theory is finitely axiomatizable (this depends on the fact that if T is finitely axiomatizable, then the set $\{\langle \alpha, \beta \rangle \mid T + \alpha \text{ interprets } T + \beta\}$ is recursively enumerable), but it is not valid if the base theory is PA. On the other hand the formula $A \rhd B \to (A \wedge \Box D) \rhd (B \wedge \Box D)$ (Montagna's principle) is valid for PA but not for the finitely axiomatizable theory GB.

This sensitivity to the base theory shows that the modal language of interpretability is considerably stronger than the modal language of provability (which does not distinguish between PA or GB).

So we have at least two different modal logics for interpretability: one for PA, called ILM (Interpretability Logic with Montagna's principle), and one for GB, called ILP.

It is still an open problem what is the interpretability logic of 'weak' theories like $I\Delta_0 + \mathrm{EXP}$, although there is a characterization in terms of Kripke-like models. For a more detailed description of these investigations we refer the reader to the papers of Visser cited in the bibliography.

Our main result, namely the completeness of ILM with respect to interpretations with base theory PA, gives a positive solution to a conjecture of Visser and makes essential use of the semantical analysis of ILM carried out by De Jongh and Veltman [5], and simplified by Visser [18]. (We need Visser's simplified semantics in our proof.)

This paper is an extended version of a preliminary manuscript [2] containing a proof of the completeness theorem. A recent preprint of Shavrukov [13] also contains a proof of the completeness theorem. The two proofs are quite different and were obtained independently from each other.

This paper is organized as follows. §2 contains some preliminary definitions and results (including Orey's theorem and some of its consequences). In §3 we describe the modal language of interpretability and state the main results. In §4 we describe, without proofs, Visser's simplified semantics of ILM. Since Visser's semantics plays an essential role in our arguments, we reproduce Visser's unpublished proof in Appendix B (with his permission). In §§5 and 6 we prove the main result, and in §7 we conclude with some open problems. Appendix A contains an exposition of De Jongh and Veltman's results (to appear in [5]) on which Visser simplified semantics is based.

§2. Preliminaries. In this section we prove Orey's theorem and some of its consequences. In particular we want to justify the use of certain model-theoretical arguments while working inside Peano arithmetic (PA). This is done by showing that these arguments are formalizable in the theory ACA_0, which is a conservative extension of PA.

2.1. Interpretability. By a 'theory' we mean an axiomatic theory formulated in first order logic (with equality). Unless otherwise stated we will only consider theories in a finite language with a recursively enumerable set of axioms.

Several notions of interpretability have been studied in the literature; the one that we are going to consider is defined in Feferman [6, p. 49], to which we refer the reader for a precise definition.

An interpretation f of a theory S into a theory T consists in an (indexed) set of formulas in the language of T which defines, in every model $\mathcal{M}$ of T, the universe, the relations (except the equality), and the (graphs of the) functions of a model $\mathcal{M}^f$ of S. (We take the equality sign as a logical symbol which is always interpreted as the identity relation.) We assume that the formula defining the universe has exactly one free variable. If $\delta(x)$ is this formula, then the underlying set of $\mathcal{M}^f$ is by definition $\{a \in \mathcal{M} \mid \mathcal{M} \vDash \delta(a)\}$.

An interpretation determines in a canonical way a map which assigns to every formula $\phi(x_1, \ldots, x_n)$ of the language of S, a formula $\phi(x_1, \ldots, x_n)^f$ of the language

of T (with the same free variables), such that for every $a_1, \ldots, a_n \in \mathcal{M}^f$,

$$\mathcal{M}^f \models \phi(a_1, \ldots, a_n) \Leftrightarrow \mathcal{M} \models \phi(a_1, \ldots, a_n)^f.$$

Since we are assuming that the equality sign is interpreted as the identity relation, we can stipulate that $(x = y)^f$ is the formula '$x = y$'.

If there is an interpretation of T_1 in T_2 we say that T_2 *interprets* T_1, or that T_1 is (relatively) *interpretable* in T_2. For example Zermelo-Fraenkel set theory (ZF) interprets Peano arithmetic (PA). If T_2 interprets T_1 via f, then clearly $T_2 \vdash \alpha^f$ for every axiom α of T_1.

A possible generalization of the above definition is given by the notion of 'n-dimensional interpretation'. This is obtained by allowing the underlying set of the interpreted model $\mathcal{M}^f$ to be a finite cartesian product of copies of $\mathcal{M}$. The typical example is the bi-dimensional interpretation of planar elementary Euclidean geometry into the theory of the real numbers. It is easy to see that for theories which have a definable pairing function (like PA), n-dimensional interpretability is equivalent to one-dimensional interpretability.

Another generalization is obtained by relaxing the requirement that '$(x = y)^f$' is '$x = y$', in order to allow the underlying set of the interpreted structure $\mathcal{M}^f$ to be a set of equivalence classes of elements from $\mathcal{M}$. However for the theories in which we are mainly interested, namely extensions of PA in the same language, one can use the induction axioms to replace each equivalence class (with respect to $=^f$) with its least representative, thus obtaining an interpretation which preserves the equality. (The same holds for any sequential theory satisfying full induction; however for theories without full induction it might actually be important to consider interpretations with equivalence classes.)

2.2. Peano arithmetic: the basic setting. We assume that the reader is familiar with the formalization of syntax in Peano arithmetic (PA) and with Gödel's incompleteness results. We recall however the basic definitions and theorems.

PA has nonlogical symbols 0, s, +, ·, for zero, successor, addition, and multiplication. The closed terms '0', 's(0)', 's(s(0))',..., are called *numerals* and '$\mathbf{n}$' denotes the nth numeral.

We denote by 'ω' the set of the natural numbers, which we identity with the first infinite ordinal number. If we consider ω together with the element zero and the operations of successor, addition, and multiplication, we obtain the standard model of PA. So the interpretation of the numeral $\mathbf{n}$ in ω is the natural number n.

Every model $\mathcal{M}$ of PA has an initial segment isomorphic to ω, which we still denote by 'ω'. Given $a \in \mathcal{M}$, we say that a is a standard element of $\mathcal{M}$ if $a \in \omega$; otherwise we say that a is a nonstandard element of $\mathcal{M}$. A model not isomorphic to ω is called a nonstandard model.

A finite sequence of symbols from the alphabet of PA is called a 'syntactical object'. So a formula, a term, or a proof, are syntactical objects. To each syntactical object t is associated bijectively (in an effective way) a natural number $\ulcorner t \urcorner$, called its 'Gödel number'.

The function which sends the natural number n to (the Gödel number of) its numeral $\mathbf{n}$ is primitive recursive, so given a model $\mathcal{M} \models$ PA this function can be

naturally prolonged to the nonstandard elements of $\mathcal{M}$ (by formalizing the definition of this primitive recursive function in PA).

DEFINITION 2.1. 1. The formula $\alpha(x_1,\dots,x_k)$ *defines* the relation R (in ω) if for all $n_1,\dots,n_k \in \omega$,

$$R(n_1,\dots,n_k) \Leftrightarrow \omega \models \alpha(n_1,\dots,n_k).$$

2. The formula $\alpha(x_1,\dots,x_k)$ *numerates* the relation R in T if for all $n_1,\dots,n_k \in \omega$,

$$R(n_1,\dots,n_k) \Leftrightarrow T \vdash \alpha(\mathbf{n_1},\dots,\mathbf{n_k}).$$

3. $\alpha(x_1,\dots,x_k)$ *binumerates* R in T if it numerates R in T and, in addition,

$$\neg R(n_1,\dots,n_k) \Leftrightarrow T \vdash \neg\alpha(\mathbf{n_1},\dots,\mathbf{n_k}).$$

THEOREM 2.2 (GÖDEL). *For every primitive recursive predicate R there is a formula $\alpha(\vec{x})$ which binumerates R in* PA.

In the above theorem the formula $\alpha(\vec{x})$ can be constructed explicitly from a primitive recursive definition of the predicate R. The formulas so obtained are called *primitive recursive formulas*, or *p.r.-formulas* for short (cf. [6, p. 53]). They have the important 'absoluteness' property that if $\mathcal{M}$ and $\mathcal{U}$ are models of PA and $\mathcal{U}$ is an end-extension of $\mathcal{M}$, then every p.r.-formula with parameters from $\mathcal{M}$ holds in $\mathcal{M}$ iff it holds in $\mathcal{U}$.

DEFINITION 2.3. The classes of formulas 'Σ_n^0' and 'Π_n^0' are defined as follows:

1. $\Sigma_0^0 = \Pi_0^0 = p.r.$

2. A formula is Σ_{n+1}^0 if it is obtained from a Π_n^0-formula by prefixing some existential quantifiers (possibly none).

3. A formula is Π_{n+1}^0 if it is obtained from a Σ_n^0-formula by prefixing some universal quantifiers.

Every recursively enumerable set can be defined (in ω) by a Σ_1^0-formula, and conversely every Σ_1^0-formula defines a recursively enumerable set.

2.3. Formalization of metamathematics in PA. The version of the diagonal lemma that we need is the following:

THEOREM 2.4 (Diagonal lemma). *For every formula $\phi(x_1,\dots,x_k,y)$, there is a formula $\alpha(x_1,\dots,x_k)$ such that* $\mathrm{PA} \vdash \forall\vec{x}(\alpha(\vec{x}) \leftrightarrow \phi(\vec{x},\ulcorner\alpha\urcorner))$.

The formula α is defined in a primitive recursive way 'relative to ϕ'; in particular if ϕ is a primitive recursive formula, then so is α (up to provable equivalence in PA).

Given an arithmetically axiomatized theory T (in the language of PA), we can associate to T, in a uniform way, a formula $\mathrm{Prf}_T(x,y)$ which formalizes the predicate 'y is a proof of x from T'. An important remark is that $\mathrm{Prf}_T(x,y)$ depends not only on the set of axioms of T, but on the way these are given, namely on the formula $\tau(x)$ defining this set. So a less ambiguous notation would be '$\mathrm{Prf}_\tau(x,y)$'. Alternatively we can think of an arithmetically axiomatized theory as coming together with the formula defining its set of axioms. The formula $\mathrm{Pr}_\tau(x)$ ('x is a theorem of τ') is defined as $\exists y\mathrm{Prf}_\tau(x,y)$, and $\mathrm{Con}(\tau)$ is $\neg\,\mathrm{Pr}_\tau(\ulcorner 0=1\urcorner)$.

It is known that the set of (the Gödel numbers of) those sentences of PA which are 'true' (in ω) is not arithmetically definable. (This is an immediate consequence of the 'diagonal lemma'.) On the other hand, for every fixed $n > 0$, there is a Σ_n^0-formula '$\mathrm{True}_{\Sigma_n^0}(x)$' which defines the set of (the Gödel numbers of) the true Σ_n^0-sentences of

PA. Moreover PA proves that $\mathrm{True}_{\Sigma_n^0}(x)$ has the properties that a truth definition should have; in particular, for every $\alpha \in \Sigma_n^0$, $\mathrm{PA} \vdash \alpha \leftrightarrow \mathrm{True}_{\Sigma_n^0}(\ulcorner \alpha \urcorner)$.

Let $n \in \omega$, $\mathscr{M} \models \mathrm{PA}$, and let $\phi \in \mathscr{M}$ be such that $\mathscr{M} \models$ '$\phi \in \Sigma_n^0$'. Since n is a standard element of $\mathscr{M}$, ϕ is (the code of) a formula with standard quantifier complexity. Note that ϕ might still contain terms of nonstandard length (as computed in $\mathscr{M}$), and therefore it is not necessarily (the code of) a standard sentence; however the partial truth definitions can be used to assign a meaning to the assertion that ϕ holds in $\mathscr{M}$.

It can be proved by induction on the complexity of α that whenever $\alpha \in \Sigma_1^0$ and $\omega \models \alpha$, then $\mathrm{PA} \vdash \alpha$. The same result can be proved also model-theoretically by observing that the Σ_1^0-formulas are preserved upwards from one model to any end-extension (so in particular from ω to any model of PA). However the syntactical proof has the advantage that it can be formalized in PA. This yields:

THEOREM 2.5 (Provable Σ_1^0-completeness of PA). PA *proves* (*a formalization of*) '*for every* ϕ, *if* ϕ *is a true* Σ_1^0*-sentence, then* $\mathrm{PA} \vdash \phi$'.

Given a formula $\sigma(x)$, we let $\sigma \restriction y$ be the formula $\sigma(x) \wedge x < y$. Similarly given a theory S, we let $S \restriction n$ be the finite subtheory of S axiomatized by all the axioms of S with Gödel number $< n$. So if $\sigma(x)$ defines a theory S, $\sigma \restriction \mathbf{n}$ defines $S \restriction n$. PA_k is defined as $\mathrm{PA} \restriction k$. If T is a finite theory, there is a canonical formula defining T, namely the formula $[T](x)$ obtained by taking the disjunction of all the formulas of the form $x = \mathbf{n}$ such that n is (the Gödel number of) an axiom of T.

An application of the partial truth definitions (and the cut elimination theorem) is that PA proves the Σ_n^0-soundness (hence the consistency) of every finite subtheory of itself (cf. [10]).

THEOREM 2.6 (Reflection). *The following is provable in* PA: *for every k and n,* PA *proves* '*for every* Σ_n^0*-sentence* ϕ, *if* $\mathrm{PA}_k \vdash \phi$, *then* ϕ *is true*'.

PROOF. The usual effective proof of Gentzen's cut elimination theorem (see [12]) is formalizable in PA. A consequence of this theorem is that if a formula χ has a proof in first order logical calculus (from no axioms), then it has a proof such that all the formulas appearing in the proof have logical complexity not bigger than the complexity of χ. Now let θ be the conjunction of the closures of all the axioms of PA_k, and let m be so big that $\theta \to \phi$ belongs to Σ_m^0. Reasoning in PA, suppose that $\mathrm{PA}_k \vdash \phi$. By the deduction theorem, $\vdash \theta \to \phi$. By the cut elimination theorem there is a derivation (from no axioms) of $\theta \to \phi$ such that all the formulas occurring in the proof belong to Σ_m^0. We can now apply the partial truth definition for Σ_m^0-formulas to show by induction on the length of the derivation that the closures of all the formulas appearing in the derivation are true. But θ is true. So ϕ must also be true. QED.

If T is a theory and ϕ is a sentence we let '$T + \phi$' denote the theory obtained by adjoining the axiom ϕ to T. As an immediate consequence of the reflection theorem, PA proves that for every k and every sentence ϕ of PA, $\mathrm{PA} + \phi \vdash \mathrm{Con}(\mathrm{PA}_k + \phi)$.

Gödel's completeness theorem says that every consistent theory has a model. From Henkin's proof of the completeness theorem it is easy to see that if T is an arithmetically definable consistent theory (in a finite language), then T has an arithmetically definable model $\mathscr{M}$, in the sense that the underlying set of the model is a definable set of integers, and the relations and functions of the model are

arithmetically definable. In fact $\mathcal{M}$ can be constructed in a recursive way from T and an oracle for $0'$, and the whole construction can be formalized in PA. Noting that a definable model is essentially an interpretation, we obtain the following formalized version of the completeness theorem (cf. [6, p. 72]).

THEOREM 2.7 (Arithmetized completeness theorem). *Let T be a theory containing* PA, *and let S be a first order theory. Suppose that $\sigma(x)$ numerates S in T and $T \vdash \mathrm{Con}(\sigma)$. Then T interprets S.*

Craig's theorem says that if a theory has a recursively enumerable set of axioms, then it is possible to find an equivalent axiomatization by a primitive recursive set of axioms. In [6, p. 63] Feferman gives the following formalized version of Craig's theorem:

THEOREM 2.8 (CRAIG). *Let $\xi(x)$ be a Σ_1^0-formula. Then there is a primitive recursive formula $\alpha(x)$ such that* $\mathrm{PA} \vdash \mathrm{Pr}_\alpha(x) \leftrightarrow \mathrm{Pr}_\xi(x)$.

The following theorem of Orey has been presented for the first time in [6, p. 80] (in a slightly weaker version).

THEOREM 2.9 (OREY). *Let T be a theory containing* PA, *and let S be a theory with a recursively enumerable set of axioms. Suppose that for every finite subtheory S' of S, $T \vdash \mathrm{Con}([S'])$. Then T interprets S.*

PROOF. By Craig's theorem we can assume that S is primitive recursively axiomatized. Let $\sigma(x)$ be a primitive recursive formula which binumerates S in PA. Let $\sigma^*(x)$ be the formula $\sigma(x) \wedge \mathrm{Con}(\sigma \restriction x + 1)$. In other words σ^* defines the union of all the consistent subtheories of σ of the form $\sigma \restriction y$. By formalizing the proof of the compactness theorem in PA, we obtain $\mathrm{PA} \vdash \mathrm{Con}(\sigma^*)$. Since $\sigma(x)$ binumerates S, the hypothesis of the theorem can be restated as $\forall k\, T \vdash \mathrm{Con}(\sigma \restriction k)$. If T is consistent this implies that $\sigma^*(x)$ numerates (actually 'binumerates') S in T (on the other hand, if T is inconsistent there is nothing to prove). Therefore, by the arithmetized completeness theorem, T interprets S. QED.

REMARK 2.10. If T is formulated in the language of PA (and contains PA), then the converse of Orey's theorem is also true: Suppose that T interprets S and S' is a finite subtheory of S. Then there must be a finite subtheory T' of T such that T' interprets S'. For finitely axiomatized theories, the interpretability relation can be formalized as a Σ_1^0-statement; thus, by the Σ_1^0-completeness of PA, $\mathrm{PA} \vdash$ 'T' interprets S''', and therefore $\mathrm{PA} \vdash \mathrm{Con}([T']) \rightarrow \mathrm{Con}([S'])$. Since T is an extension of PA in the same language, by the reflection theorem $T \vdash \mathrm{Con}([T'])$. Thus $T \vdash \mathrm{Con}([S'])$.

REMARK 2.11. Orey's theorem and its converse continue to hold for every sequential theory satisfying 'full induction' (see the Introduction). Any such theory admits partial truth definitions and proves the cut elimination theorem; hence it is not finitely axiomatizable (by a version of the reflection theorem).

2.4. A characterization of interpretability over PA.

THEOREM 2.12. *Let $\mathcal{M}$ be a model of* PA, *and let f be an interpretation of* PA *into the theory of $\mathcal{M}$. This means that the interpreted structure $\mathcal{M}^f$ is a model of* PA. *We claim that $\mathcal{M}$ can be embedded as an initial segment of $\mathcal{M}^f$. (Recall that we are considering interpretations which preserve the equality.)*

PROOF. Let $G(x, y)$ be the formula of PA which formalizes the following: there is a finite sequence $u = \langle u_0, \ldots, u_x \rangle$ such that $(u_0 = 0)^f$, $u_x = y$, and, for all $i < x$,

$(u_{i+1} = s(u_i))^f$. By the induction axioms it follows that for every $x \in \mathcal{M}$ there is a (unique) y in $\mathcal{M}^f$ such that $\mathcal{M} \models G(x, y)$. Let $g(x)$ be the unique element $y \in \mathcal{M}^f$ such that $\mathcal{M} \models G(x, y)$. Then $\mathcal{M}^f \models g(x + 1) = g(x) + 1$, and g is an embedding of $\mathcal{M}$ as a submodel of $\mathcal{M}^f$. (It is easy to see that the embedding g preserves $+$ and $\times$.) The sentence $\forall x, u(u < x + 1 \to u < x \vee u = x)$ is a theorem of PA, and therefore it is true in $\mathcal{M}^f$. Therefore $\mathcal{M} \models (u < g(x + 1) \leftrightarrow u < g(x) \vee u = g(x))^f$. It follows by induction on $x \in \mathcal{M}$ that for every $u \in \mathcal{M}$ satisfying $\mathcal{M} \models (u < g(x))^f$, there is some $y < x$, in $\mathcal{M}$, such that $\mathcal{M} \models (u = g(y))^f$. This means that g embeds $\mathcal{M}$ as an initial segment of $\mathcal{M}^f$. QED.

THEOREM 2.13. *Let α and β be sentences of* PA. *Then* PA $+ \alpha$ *interprets* PA $+ \beta$ *iff every model of* PA $+ \alpha$ *has an end-extension which is a model of* PA $+ \beta$.

PROOF. The '$\Rightarrow$' part follows immediately from the previous theorem. Conversely, if PA $+ \alpha$ does not interpret PA $+ \beta$, by Orey's theorem there is some k such that PA $+ \alpha \nvdash \mathrm{Con}(\mathrm{PA}_k + \beta)$. Take a model $\mathcal{M}$ of PA $+ \alpha$ in which $\neg\mathrm{Con}(\mathrm{PA}_k + \beta)$ holds. Being a Σ^0_1-assertion, $\neg\mathrm{Con}(\mathrm{PA}_k + \beta)$ must then hold in any end-extension $\mathcal{U}$ of $\mathcal{M}$. But then by the reflection theorem $\mathcal{U}$ cannot be a model of PA $+ \beta$. QED.

2.5. Doing model theory inside PA. The theorem just proved is not directly formalizable in PA, since it speaks about models and a model is in general an infinite object. However it is formalizable and provable in the conservative extension of PA known as ACA_0.

DEFINITION 2.14. ACA_0 is a first order theory formulated in a language properly containing the language of PA. It has two sorts of variables, set variables X, $Y, \ldots$, and numerical variables x, $y, \ldots$ (this can be translated into the usual formalization of first order logic with just one sort of variables by adding a unary predicate for 'x is a set'). We also have a binary relation '$x \in X$' whose intended meaning is that the number x is an element of the set X. The axioms of ACA_0 are the following:

1. All the axioms of PA except the induction scheme.
2. (Induction) $\forall X(0 \in X \wedge \forall x(x \in X \to x + 1 \in X) \to \forall x(x \in X))$.
3. (Extensionality) $\forall x(x \in X \leftrightarrow x \in Y) \to X = Y$.
4. (Arithmetical comprehension) Let $\phi(x)$ be a formula containing no bound set variables ($\phi(x)$ might contain several free variables both of the number sort and of the set sort), and let X be a set variable not occurring in ϕ. Then there is an axiom which asserts the universal closure of $\exists X \forall x(x \in X \leftrightarrow \phi(x))$.

Every model $\mathcal{M}$ of PA can be expanded to a model of ACA_0 by interpreting the set variables as ranging over the definable subset of $\mathcal{M}$ (with parameters), and by interpreting '$\in$' as the usual membership relation. This construction also shows that ACA_0 is conservative over PA; that is, every sentence in the language of PA which is provable in ACA_0 is already provable in PA.

When working inside ACA_0 we stipulate that a 'model' consists of an underlying set, together with the various relations and functions of the model; and, in addition to that, we also require that a model contains the satisfaction relation for its structure. We also need to take care of the following minor point: in ACA_0 there is a set which is not properly included in any other set, namely $\{x \mid x = x\}$. Consequently embeddings of models cannot always be replaced by inclusions (we thank Professor P. Hájek for this remark). To avoid this problem we stipulate that whenever we say

"end-extension" we really mean "isomorphic to an end-extension". With this proviso it is easy to verify that the proof of Theorem 2.13 can be formalized in ACA_0:

THEOREM 2.15. ACA_0 *proves the following*: PA + α *interprets* PA + β *iff every model of* PA + α *has an end-extension which is a model of* PA + β.

It is also well known that in ACA_0 one can carry out Henkin's proof of Gödel's completeness theorem.

Now let $\mathcal{M}$ be a model of PA, and suppose that $\forall k\ \mathcal{M} \models \mathrm{Con}(\mathrm{PA}_k + \phi)$. By Orey's theorem the theory of $\mathcal{M}$ interprets the theory PA + ϕ, and therefore there is an end-extension $\mathcal{U}$ of $\mathcal{M}$ (namely the interpreted structure) which is a model of PA + ϕ. It is not difficult to verify that this construction can be formalized in ACA_0 and can be extended to the case in which ϕ is allowed to contain nonstandard terms (as computed in $\mathcal{M}$). So we get:

THEOREM 2.16. ACA_0 *proves the following. Let* $\mathcal{M}$ *be a model of* PA, *let* $\phi(x)$ *be a formula, and let m be an element of* $\mathcal{M}$. *Suppose that* $\forall k\ \mathcal{M} \models \mathrm{Con}(\mathrm{PA}_k + \phi(\mathbf{m}))$. *Then there is an end-extension* $\mathcal{U}$ *of* $\mathcal{M}$ *such that* $\mathcal{U} \models \mathrm{PA} + \phi(m)$.

Note that the (possibly nonstandard) term '**m**' has the same meaning in $\mathcal{M}$ and in $\mathcal{U}$. This is so because the evaluation function **eval** that sends the (Gödel number of the) term '**m**' to its denotation 'm' is primitive recursive, and therefore gives the same output in $\mathcal{M}$ and in $\mathcal{U}$. (The same would be true if we start with any closed term t rather than the numeral **m**.)

§3. Modal logic for interpretability.

DEFINITION 3.1 (The modal language). The *modal language of interpretability*, $\mathscr{L}(\triangleright)$, has an infinite set of propositional variables $v_0, v_1, \ldots$, a propositional constant $\bot$ (denoting falsehood), the binary propositional connectives '$\wedge$' (conjunction) and '$\neg$' (negation), a binary propositional operator '$\triangleright$' (for relative interpretability over PA), and a unary operator '$\Box$' (for provability in PA). The formulas are defined inductively as follows: 1) a propositional variable is a formula; 2) $\bot$ is a formula; 3) if A and B are formulas, then so are $\neg A$, $A \triangleright B$, $A \wedge B$, and $\Box A$. We let $\Diamond A$ be $\neg \Box \neg A$, and we define the other propositional connectives in terms of $\neg$ and $\wedge$ in the usual way. The *modal language of provability*, $\mathscr{L}(\Box)$, is the fragment of $\mathscr{L}(\triangleright)$ obtained by omitting the symbol '$\triangleright$'.

The intended meaning of '$A \triangleright B$' is 'PA + A interprets PA + B', and the intended meaning of '$\Box A$' is 'PA $\vdash A$'. This is made precise in the following definition:

DEFINITION 3.2 (Interpreting $\mathscr{L}(\triangleright)$ in the language of PA). An *interpretation* f of $\mathscr{L}(\triangleright)$ in the language of PA is a map which assigns to every propositional variable v a sentence v^f of PA. We extend f to all the modal formulas as follows:

1. $(A \triangleright B)^f$ is given by a formalization of 'PA + A^f interprets PA + B^f'.
2. $(\Box A)^f$ is a formalization of 'PA $\vdash A^f$'.
3. $(\bot)^f$ is the PA-sentence '0 = 1'.
4. f commutes with propositional connectives, so $(A \wedge B)^f$ is $A^f \wedge B^f$ and $(\neg A)^f$ is $\neg(A^f)$.

Note that by Orey's theorem and its converse 'PA + ψ interprets PA + χ' is equivalent to the Π^0_2-assertion '$\forall k$ PA + $\psi \vdash \mathrm{Con}(\mathrm{PA}_k + \chi)$'. (An analysis of the proof shows that this equivalence is provable in PA.)

DEFINITION 3.3. We say that a modal formula A is PA-*valid* if, for every interpretation f, A^f is a theorem of PA. We say that A is ω-*valid* if, for every interpretation f, A^f is true (namely, it holds in the standard model ω).

Note that A is PA-valid if and only if $\Box A$ is ω-valid. Also note that every PA-valid formula is also ω-valid.

It is worth mentioning that provability is definable in terms of interpretability; namely, the formula $\Box A \leftrightarrow (\neg A) \rhd \bot$ is PA-valid. If we wished we could have used this equivalence to define $\Box$ in terms of $\rhd$, omitting $\Box$ from the modal language altogether.

DEFINITION 3.4. The modal theory L is formulated in the language of provability $\mathscr{L}(\Box)$ and is axiomatized by all the tautologies (including those containing the $\Box$-operator) plus the following axiom schemes:

1. $\Box(A \to B) \to (\Box A \to \Box B)$,
2. $\Box(\Box A \to A) \to \Box A$,
3. $\Box A \to \Box\Box A$.

The rules of inference are: 1) if $\vdash A \to B$ and $\vdash A$, then $\vdash B$ (modus ponens); 2) if $\vdash A$, then $\vdash \Box A$ (necessitation).

DEFINITION 3.5. The modal theory L^ω is formulated in $\mathscr{L}(\Box)$ and is axiomatized by all the theorems of L plus all the instances of the scheme $\Box A \to A$. The only rule of inference is modus ponens.

In [16] Solovay gave the following axiomatizations for the valid formulas of the language of provability $\mathscr{L}(\Box)$:

THEOREM 3.6 (SOLOVAY). *Let $A \in \mathscr{L}(\Box)$. Then A is* PA-*valid iff $L \vdash A$, and A is ω-valid iff $L^\omega \vdash A$.*

In Solovay's original paper the theory L is called 'G'. We use here the terminology of [5].

In [18] Visser defined the modal theory ILM (there called $\mathrm{IL}_{\mathrm{ERA}}$) and conjectured that ILM axiomatizes the PA-valid formulas of the modal language of interpretability. The name 'ILM' stands for 'Interpretability Logic with Montagna's principle', where Montagna's principle is the axiom scheme $A \rhd B \to (A \wedge \Box D) \rhd (B \wedge \Box D)$.

DEFINITION 3.7 (The theory ILM). The modal theory ILM is formulated in the language $\mathscr{L}(\rhd)$ and is axiomatized by the axiom schemes listed below plus all the tautologies.

The axioms. 1. $\Box(A \to B) \to (\Box A \to \Box B)$.
2. $\Box(\Box A \to A) \to \Box A$.
3. $\Box A \to \Box\Box A$.
4. $\Box(A \to B) \to A \rhd B$.
5. $(A \rhd B \wedge B \rhd C) \to A \rhd C$.
6. $(A \rhd C \wedge B \rhd C) \to (A \vee B) \rhd C$.
7. $A \rhd B \to (\Diamond A \to \Diamond B)$.
8. $\Diamond A \rhd A$.
9. $A \rhd B \to (A \wedge \Box D) \rhd (B \wedge \Box D)$.

The rules of inference are modus ponens and necessitation.

In §4 we will give a positive solution to Visser's conjecture by showing that the PA-valid formulas are exactly the theorems of ILM:

THEOREM 3.8 (Arithmetical completeness of ILM). *The* PA*-valid formulas of* $\mathscr{L}(\triangleright)$ *are exactly the theorems of* ILM.

The 'soundness' part was already known, and amounts to noticing that all the axioms are PA-valid and the rules of inference preserve PA-validity. Axiom 1, for example, is a formalization of the deduction theorem. Axiom 2 is a formalization of Löb's theorem (a statement essentially equivalent to a formalization of Gödel's second incompleteness theorem). To prove it, work in PA and assume that $(\Box A)^f$ fails, that is, $\mathrm{PA} \nvdash A^f$; then $\mathrm{PA} + \neg A^f$ is a consistent theory and therefore does not prove its own consistency: $\mathrm{PA} + \neg A^f \nvdash \mathrm{Con}(\mathrm{PA} + \neg A^f)$. But this is equivalent to $\mathrm{PA} \nvdash (\text{'PA} \vdash A^f\text{'} \rightarrow A^f)$, namely $\neg(\Box(\Box A \rightarrow A))^f$. Axiom 3 is an expression of the provable Σ_1^0-completeness of PA: in fact, reasoning in PA, if the Σ_1^0-assertion $(\Box A)^f$ holds, then it must be provable, namely $(\Box\Box A)^f$. Axiom 7 says that relative interpretability yields relative consistency results. Axiom 8 is the arithmetized completeness theorem: PA plus the assertion that a given theory is consistent interprets the given theory. Axiom 9 (Montagna's principle) is a consequence of Orey's theorem; an easy way to prove it is to make use of Theorem 2.13: $\mathrm{PA} + \alpha$ interprets $\mathrm{PA} + \beta$ if and only if every model of $\mathrm{PA} + \alpha$ has an end-extension which is a model of $\mathrm{PA} + \beta$; Montagna's principle now follows immediately from the observation that the Σ_1^0-sentence $(\Box D)^f$ must be preserved upwards under end-extensions. Theorem 2.13 can also be used to verify axiom 6 (although there is a more elementary proof). The other axioms are easy to verify.

To obtain a characterization of the ω-valid formulas we have to add to ILM some axioms that are ω-valid but not PA-valid. The simplest example is the scheme $\Box A \rightarrow A$ asserting the soundness of PA. This yields the theory ILM^ω defined as follows:

DEFINITION 3.9 (The theory ILM^ω). The axioms of the theory ILM^ω are all the theorems of ILM plus all the instances of the axiom scheme $\Box A \rightarrow A$. The only rule of inference of ILM^ω is modus ponens.

In §6 we will prove the following:

THEOREM 3.10 (Arithmetical completeness of ILM^ω). *The* ω*-valid formulas are exactly the theorems of* ILM^ω.

De Jongh and Veltman [5] proved that the theory ILM is decidable. ILM^ω is also decidable, since the proof of our results will show that ILM^ω can be recursively reduced to ILM (Theorem 6.5). So we have a decision procedure for the classes of the PA-valid and ω-valid formulas. Moreover, if a formula is not PA-valid (or ω-valid), we will be able to construct an explicit counterexample.

3.1. Plan of the proof. The main ingredient in the proof of our results is a systematic procedure to build arithmetical interpretations to provide the needed counterexamples. The general plan is the following: if $\mathrm{ILM} \vdash A$, then we have already remarked that A is PA-valid. On the other hand, if $\mathrm{ILM} \nvdash A$, then by the results of [5] and [18] (which we present below) there is a Kripke-like model $\mathbf{V}$ of ILM in which A fails; at this point we will be able to apply our procedure to transform $\mathbf{V}$ into an arithmetical interpretation 'I', the *induced interpretation*, such that $\mathrm{PA} \nvdash A^I$. A similar construction will show that if $\mathrm{ILM}^\omega \nvdash A$, then A is not ω-valid.

§4. Semantics for interpretability logic. The situation is the following: to prove the decidability of ILM de Jongh and Veltman showed that ILM is complete with respect to a certain class of 'Kripke-like' finite models which they called 'ILM-models'. The finiteness of the models is what establishes the decidability. We will give an exposition of their proof in Appendix A. In [18] Visser showed that ILM is also complete with respect to a different class of models, the 'simplified ILM-models' (which, although simpler, are no longer finite). Visser's simplified models are obtained by expanding each ILM-model into an equivalent simplified ILM-model in a certain primitive recursive manner (the elements of the induced simplified model are finite sequences of elements of the ILM-model, as explained in Appendix B). For the purposes of our work we found it more convenient to make use of the simplified ILM-models, which we define in this section. As a preliminary step we begin by defining the models for the logic L of provability (those employed by Solovay for the proof of Theorem 3.6).

4.1. L-models.

DEFINITION 4.1. An *L-frame* is a tuple $\langle V, R, b\rangle$, where:

1. V is a nonempty set;
2. R is transitive binary relation on V which is reverse well-founded, that is, there is no infinite sequence $x_1 R x_2 R x_3 \cdots$; and
3. b is an element of V, called the *root*, such that $\forall x \in V(x = b \vee bRx)$.

Note that the reverse well-foundedness of R implies in particular that R is antireflexive: $\forall x \in V(\neg xRx)$. Sometimes we will call the elements of V 'worlds' or 'nodes'.

DEFINITION 4.2. An *L-model* is given by an L-frame together with a forcing relation '$\Vdash$', included in $V \times \mathscr{L}(\Box)$, which commutes with Boolean connectives and satisfies

$$u \Vdash \Box A \Leftrightarrow \forall v(uRv \Rightarrow v \Vdash A).$$

Given an L-model $\mathbf{V} = \langle V, R, b, \Vdash\rangle$, we define $\mathbf{V} \vDash A$ iff $\forall v \in V(v \Vdash A)$. It is easy to verify that L is sound with respect to the L-models; that is, if $L \vdash A$, then for every L-model $\mathbf{V}$, $\mathbf{V} \vDash A$. It is also known that L is complete with respect to the class of all the finite L-models (hence L is decidable):

THEOREM 4.3 (Modal completeness of L). *If $L \nvdash A$, then there is a finite L-model $\mathbf{V}$, with root b say, such that $b \Vdash \neg A$.*

For a proof of this theorem we refer the reader to [4] or [16].

4.2. Simplified ILM-models.

DEFINITION 4.4. A *simplified* ILM-*frame* is an L-frame $\langle V, R, b\rangle$ together with an additional relation S with the following properties:

1. S is a transitive, reflexive, binary relation on V such that $R \subseteq S$.
2. $\forall x, y, z \in V(xSyRz \Rightarrow xRz)$.

DEFINITION 4.5. A *simplified* ILM-*model* is given by a simplified ILM-frame $\langle V, R, S, b\rangle$ together with a *forcing* relation '$\Vdash$', included in $V \times \mathscr{L}(\triangleright)$, which commutes with the Boolean connectives and satisfies:

1. $u \Vdash A \triangleright B \Leftrightarrow \forall v(uRv \wedge v \Vdash A \Rightarrow \exists w(uRw \wedge vSw \wedge w \Vdash B))$;
2. $u \Vdash \Box A \Leftrightarrow \forall v(uRv \Rightarrow v \Vdash A)$.

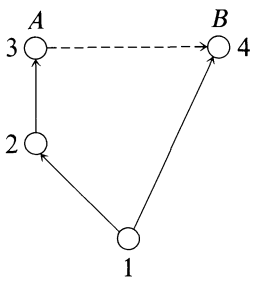

FIGURE 1

Given a simplified ILM-model $\mathbf{V} = \langle V, R, S, b, \Vdash \rangle$, we define $\mathbf{V} \models A$ iff $\forall v \in V(v \Vdash A)$. It is routine to verify that ILM is sound with respect to the simplified ILM-models, that is:

PROPOSITION 4.6 (Modal soundness of ILM). *If* ILM $\vdash A$, *then for every simplified* ILM*-model* $\mathbf{V}$, $\mathbf{V} \models A$.

EXAMPLE 4.7. In Figure 1 the dashed arrow represent an S-relation and the full arrows represent R-relations.

The figure is supposed to represent the minimal ILM-model satisfying the conditions $1R2$, $2R3$, $1R4$, and $3S4$. (By the transitivity of R there is an R-relation between 1 and 3 which is not indicated in the figure. Also, we must have $2S4$.) The figure indicates that the atomic formula A holds at the node 3 and nowhere else, and the atomic formula B holds only at 4. The reader can verify that $1 \Vdash A \rhd B$, but $2 \nVdash A \rhd B$. Thus $1 \Vdash A \rhd B \wedge \neg \Box(A \rhd B)$, witnessing the fact that ILM $\nvdash A \rhd B \rightarrow \Box(A \rhd B)$.

The version of the completeness theorem that we present below has been proved by Visser (in the technical report [18]) via a reduction to the completeness theorem of de Jongh and Veltman. We reproduce Visser's unpublished proof in Appendix B (with his permission).

THEOREM 4.8 (Completeness of ILM with respect to simplified models). *If* ILM $\nvdash A$, *then there is a countable simplified* ILM*-model* $\langle V, R, S, b, \Vdash \rangle$ *such that* $b \Vdash \neg A$.

4.3. Recursiveness of Visser's simplified models. The simplified models needed for the completeness theorem of ILM cannot be taken to be finite. (In a private communication Visser constructed a modal formula that has an infinite simplified model but not a finite one.) However, we will show that they can be taken to be 'primitive recursive', which is almost as good for proving that certain relations are absolute for models of PA.

DEFINITION 4.9. We say that $\langle V, R, S, b, \Vdash \rangle$ is a *primitive recursive* ILM*-model* if it is a simplified ILM-model with the following additional properties:

1. V is a primitive recursive set of natural numbers (in the sense that the characteristic function of V is primitive recursive).

2. R and S are primitive recursive binary relations on V.

3. The relation $\{\langle x, A \rangle \mid x \Vdash A\}$ is primitive recursive (here A ranges over all the modal formulas and is not restricted to be atomic).

4. Let $Q = \{\langle x, y, B\rangle \mid \exists z\colon xRz \wedge ySz \wedge z \Vdash B\}$. Then Q is a primitive recursive relation.

5. The following strengthening of the well-foundedness property of R holds: there is a natural number k such that every (finite) sequence $x_1 R x_2 R \cdots$ has length $\leq k$.

Clearly any finite ILM-model is primitive recursive.

In the following when we refer to a primitive recursive ILM-model $\mathbf{V}$ we will always assume that $\mathbf{V}$ is given together with an explicit primitive recursive definition of V, R, S, the forcing relation, and the predicate Q. By abuse of terminology we will denote by 'V', 'R', 'S', '$\Vdash$', and 'Q', both the relations and the formulas of PA corresponding to the respective primitive recursive definitions.

DEFINITION 4.10. We say that $\mathbf{V} = \langle V, R, S, b, \Vdash\rangle$ is a *provably primitive recursive* ILM-*model* if $\mathbf{V}$ is a primitive recursive ILM-model and there is a primitive recursive formula $Q(x, y, B)$ such that

$$\mathrm{PA} \vdash Q(x, y, B) \leftrightarrow \exists z(xRz \wedge ySz \wedge z \Vdash B),$$

and PA proves that $\langle V, R, S, b, \Vdash\rangle$ is a simplified ILM-model satisfying the given strengthened version of the well-foundedness property of R.

A careful analysis of Visser's proof of Theorem 4.8 shows that we have in fact the following strengthening:

THEOREM 4.11. *If* $\mathrm{ILM} \nvdash A$, *then there is a provably primitive recursive* ILM-*model* $\mathbf{V}$, *with root* b *say, such that* $b \Vdash A$.

The only part that does not follow trivially from Visser's proof is that the relation

$$Q = \{\langle x, y, B\rangle \mid \exists z\colon xRz \wedge ySz \wedge z \Vdash B\}$$

is primitive recursive. This is shown in Appendix B.

Note that a finite simplified ILM-model $\langle V, R, S, b, \Vdash\rangle$ is always provably primitive recursive.

§5. Arithmetical completeness of ILM. In this section we prove that the PA-valid formulas of $\mathscr{L}(\triangleright)$ are exactly the theorems of ILM.

5.1. The induced interpretation. Fix a provably primitive recursive ILM-model $\mathbf{V} = \langle V, R, S, b, \Vdash\rangle$. Without loss of generality we assume that $b = 1$ and $0 \notin V$.

DEFINITION 5.1 (Adjoining a new root). We extend R and S to the set $V \cup \{0\}$ by setting $0Rx$ for all $x \in V$ and $0Sx$ for all $x \in V \cup \{0\}$. We also extend the forcing relation to the node 0 by defining $0 \Vdash A \Leftrightarrow 1 \Vdash A$ for every atomic formula A. This gives a new simplified ILM-model with underlying set $V \cup \{0\}$ and root 0. It is easy to verify that the two models agree on their common domain in the sense that for all $x \in V$ and for every modal formula A (not necessarily atomic), if $x \Vdash A$ in one model, then $x \Vdash A$ in the other model.

According to the general plan of the proof (see §3.1) we need to define an interpretation 'I' of $\mathscr{L}(\triangleright)$ into the language of PA, the *induced interpretation*, such that whenever $1 \Vdash A$, then $\mathrm{PA} \vdash A^I$. We will define in PA a constant L (depending on $\mathbf{V}$) and then define I in terms of L as follows:

DEFINITION 5.2. For A atomic, A^I is the sentence of PA expressing the following:

$$\exists x \in V \cup \{0\}\colon L = x \wedge x \Vdash A.$$

L will be defined as the limit of a primitive recursive function $F: \omega \to V \cup \{0\}$ in such a way that the following properties hold:

5.2. Properties of the function F. F is a primitive recursive function and L is a definable constant of PA such that PA proves:

(R) For all x, y in $V \cup \{0\}$, if $L = x$ and xRy, then $\mathrm{Con}(\mathrm{PA} + L = \mathbf{y})$.

$(\neg R)$ For all $x \in V$, if $L = x$, then $\neg \mathrm{Con}(\mathrm{PA} + \exists y: L = y \wedge \neg xRy)$.

(S) For all $x \in V \cup \{0\}$ if $L = x$, then for all k PA proves that for all $y, z \in V \cup \{0\}$, if $L = y$, $\mathbf{x}Rz$ *and* ySz, then $\mathrm{Con}(\mathrm{PA}_k + L = \mathbf{z})$.

$(\neg S)$ L is the limit of the function $F: \omega \to V \cup \{0\}$, and for all n, m if $n \leq m$, then $F(n)SF(m)$.

In the following we will omit the boldface notation for numerals when the meaning is clear from the context. So we will write $\mathrm{Con}(\mathrm{PA} + L = y)$ rather than $\mathrm{Con}(\mathrm{PA} + L = \mathbf{y})$.

REMARK 5.3. Since $\forall x \in V(\neg xRx)$, property $(\neg R)$ implies that, for all $x \in V$, $\mathrm{PA} + L = x \vdash \neg \mathrm{Con}(\mathrm{PA} + L = x)$. It follows that in the standard model $L = 0$. Thus, by property (R), for all $x \in V$ $\mathrm{PA} + L = x$ is consistent (as $0Rx$). Of course $\mathrm{PA} + L = 0$ is also consistent because $L = 0$ holds in the standard model.

REMARK 5.4. Note that if V is finite, then the first three properties of F are equivalent to the assertion that, for every $x, y, z \in V \cup \{0\}$,

1. if xRy, then $\mathrm{PA} + L = x \vdash \mathrm{Con}(\mathrm{PA} + L = y)$;
2. if $x \neq 0$ and $\neg(xRy)$, then $\mathrm{PA} + L = x \vdash \neg \mathrm{Con}(\mathrm{PA} + L = y)$;
3. if xRz and ySz, then $\mathrm{PA} + L = x$ proves that $\mathrm{PA} + L = y$ interprets $\mathrm{PA} + L = z$.

The equivalence can be shown by observing that for V finite, the assertion '$\mathrm{PA} \vdash \forall x \in V\ \psi$' is equivalent to '$\forall x \in V\ \mathrm{PA} \vdash \psi$', and therefore we can bring all the quantifiers restricted to V outside of the scope of the provability predicate. (For the third property we also need to use Orey's theorem.)

5.3. Proof that the properties of F imply the main result. Assume that for every provably primitive recursive ILM-model $\mathbf{V}$ we can define F (and L) possessing the above properties. We will show how to derive the main result from this assumption.

DEFINITION 5.5. Let A be a modal formula. Working in PA, we say that I is *faithful on* A if for all $x \in V$ we have

1. if $x \Vdash A$ and $L = x$, then A^I, and
2. if $x \Vdash \neg A$ and $L = x$, then $\neg A^I$.

It is clear that I is faithful on atomic formulas. We will show that for all C, PA proves that I is faithful on C. Once we have done this we can prove the main result as follows: if $\mathrm{ILM} \nvdash A$, then there is a provably primitive recursive ILM-model $\mathbf{V}$ with root $b \in V$, such that $b \Vdash \neg A$. Thus $\mathrm{PA} + L = b \vdash \neg A^I$. But $\mathrm{PA} + L = b$ is consistent. So $\mathrm{PA} \nvdash A^I$, and we are done.

In the proof of the following lemma we will use Theorem 2.15: $\mathrm{PA} + \alpha$ interprets $\mathrm{PA} + \beta$ iff every model of $\mathrm{PA} + \alpha$ has an end-extension which is a model of $\mathrm{PA} + \beta$. We will invoke this theorem while working in PA; this use of models in PA is justified by the fact that the above characterization of interpretability is provable in the theory ACA_0, which is a conservative extension of PA. We will also make use of Theorem 2.16 (with the same justification).

LEMMA 5.6. *For all C,* PA *proves that I is faithful on C.*

PROOF. The proof is by induction on the complexity of C. If C is atomic, the thesis is clear. If C is a Boolean combination of simpler formulas the thesis follows immediately from the induction hypothesis. Since $\Box A$ is ILM-equivalent to $(\neg A) \rhd \bot$, a little thought will convince the reader that we only need to consider the case when $C = A \rhd B$. So assume by the induction hypothesis that $\mathrm{PA} \vdash$ 'I is faithful on A' and $\mathrm{PA} \vdash$ 'I is faithful on B'. We need to prove the two clauses of the definition of 'faithful'.

Part 1. Work in PA. Suppose that $x \in V$, $x \Vdash A \rhd B$, and $L = x$. We must prove that $\mathrm{PA} + A^I$ interprets $\mathrm{PA} + B^I$. Deny this. Then there is a model $\mathscr{Y}$ of $\mathrm{PA} + A^I$ which does not have any end-extension $\mathscr{Z}$ which is a model of $\mathrm{PA} + B^I$.

Claim 1. *There is an element $y \in \mathscr{Y}$ such that $\mathscr{Y} \models xRy \wedge y \Vdash A$.*

To prove the claim we take y to be the unique element of $\mathscr{Y}$ such that $\mathscr{Y} \models L = y$. Since $x \in V$, $L = x$, and $\mathscr{Y} \models L = y$, by property $(\neg R)$ we must have xRy. To show that $\mathscr{Y} \models y \Vdash A$, we use one of the induction hypothesis: $\mathrm{PA} \vdash$ 'I is faithful on A'. (There is a subtle point here: we are now working in PA but the induction hypothesis was assumed outside of PA. However, the induction hypothesis is a Σ^0_1-assertion, namely the assertion that something is provable, so it holds inside PA as well.) Since $\mathscr{Y}$ is a model of PA, $\mathscr{Y} \models$ 'I is faithful on A'. Moreover $\mathscr{Y} \models L = y \wedge A^I$, so we must have $\mathscr{Y} \models (y \Vdash A)$. Thus the claim is proved.

Now notice that the assumption '$x \Vdash A \rhd B$', being a Σ^0_1-assertion, must hold in $\mathscr{Y}$. So we can apply the definition of '$x \Vdash A \rhd B$' inside $\mathscr{Y}$ to conclude, using the claim, that $\exists z \in \mathscr{Y}$ such that $\mathscr{Y} \models xRz \wedge ySz \wedge z \Vdash B$. Since $L = x$ and $\mathscr{Y} \models L = y \wedge xRz \wedge ySz$, by property (S) we must have $\forall k\, \mathscr{Y} \models \mathrm{Con}(\mathrm{PA}_k + L = z)$. So there is an end-extension $\mathscr{Z}$ of $\mathscr{Y}$ such that $\mathscr{Z} \models \mathrm{PA} + L = z$ (by Theorem 2.16). To reach a contradiction we will show that $\mathscr{Z} \models \mathrm{PA} + B^I$. Since '$z \Vdash B$' is a Σ^0_1-assertion, being true in $\mathscr{Y}$ it must hold in its end-extension $\mathscr{Z}$. By the induction hypothesis $\mathrm{PA} \vdash$ 'I is faithful on B', so $\mathscr{Z} \models$ 'I is faithful on B'; therefore from the fact that $\mathscr{Z} \models L = z \wedge z \Vdash B$ we can conclude that $\mathscr{Z} \models B^I$, which is the desired contradiction.

Part 2. Work in PA. Assume $x \in V$, $x \Vdash \neg(A \rhd B)$, and $L = x$. So there is some $y \in V$ such that

$$xRy \wedge y \Vdash A \wedge \neg \exists w \in V (xRw \wedge ySw \wedge w \Vdash B).$$

We must prove that $\mathrm{PA} + A^I$ does not interpret $\mathrm{PA} + B^I$. To prove this it is enough to find a model $\mathscr{Y}$ of $\mathrm{PA} + A^I$ which does not have any end-extension $\mathscr{Z}$ which is a model of $\mathrm{PA} + B^I$. We claim that any model $\mathscr{Y}$ of $\mathrm{PA} + L = y$ has the desired properties. First of all, notice that since $L = x$ and xRy, it follows that $\mathrm{PA} + L = y$ is consistent (by property (R)); so $\mathrm{PA} + L = y$ indeed has a model, $\mathscr{Y}$ say. We need to show that $\mathscr{Y} \models \mathrm{PA} + A^I$. Since '$y \Vdash A$' is a true Σ^0_1-assertion, it must hold inside the model $\mathscr{Y}$ as well. By the induction hypothesis, $\mathrm{PA} \vdash$ 'I is faithful on A'; so $\mathscr{Y} \models$ 'I is faithful on A'. Therefore from the fact that $\mathscr{Y} \models L = y \wedge y \Vdash A$ we can conclude that $\mathscr{Y} \models A^I$. Now suppose by contradiction that $\mathscr{Y}$ has an end-extension $\mathscr{Z}$ which is a model of $\mathrm{PA} + B^I$. We will prove:

Claim 2. *Let $z \in \mathscr{Z}$ be such that $\mathscr{Z} \models L = z$. Then $\mathscr{Z} \models z \in V \wedge xRz \wedge ySz$.*

To prove this, first note that $\mathscr{Z}$ must satisfy all the Σ^0_1-formulas which are true in $\mathscr{Y}$, so $\mathscr{Z} \models y \in \mathrm{Range}(F)$; hence, by property $(\neg S)$, $\mathscr{Z} \models ySz$. From $xRySz$ it follows in particular that $z \neq 0$; hence $z \in V$. (Notice that we have not yet used the

hypothesis $x \in V$.) Since $L = x$, $x \in V$, and $\mathscr{Z} \models \mathrm{PA} + L = z$, by property $(\neg R)$ we must have $\mathscr{Z} \models xRz$. This proves the claim.

By our choice of y we have $\neg \exists w \in V(xRw \wedge ySw \wedge w \Vdash B)$. In the ILM-model that we are considering this last assertion is primitive recursive, so it must be satisfied in the model $\mathscr{Z}$. In particular, taking $w = z$, we find that $\mathscr{Z} \models (z \Vdash \neg B)$. By induction hypothesis $\mathrm{PA} \vdash$ 'I is faithful on B'. So $\mathscr{Z} \models \neg B^I$. Absurd. QED.

5.4. The definition of F. To finish the proof of the main result we have to define F. The definition will involve the Gödel number of the formula defining its limit L. The apparent circularity is handled by the diagonal lemma in the usual way. So let $L = \lim F$ if F has a limit, $L = 0$ otherwise. Before defining F we need some auxiliary definitions:

DEFINITION 5.7. Let $x \in V$. We define the *rank* of x at stage n, $\mathrm{rank}(x, n)$, as the least number $i \leq n$ such that there is a proof of Gödel number $\leq n$ of $L \neq x$ from PA_i. If i does not exist we define $\mathrm{rank}(x, n)$ to be the ordinal number ω.

Intuitively, the smaller $\mathrm{rank}(x, n)$ is, the more inconsistent is the fact that $L = x$. It is clear that after a suitable coding of the ordinals $\leq \omega$ the rank function can be coded as a primitive recursive function, and its definition can be formalized in PA. Note that for a fixed x, $\mathrm{rank}(x, n)$ is a nonincreasing function of n (with respect to the natural ordering of the ordinals $\leq \omega$), and its limit is either ω (if $\mathrm{PA} + L = x$ is consistent) or the least i such that $\mathrm{PA}_i + L = x$ is inconsistent. By the reflection theorem this i will be nonstandard in any model of $\mathrm{PA} + L = x$; more precisely, we have: let $\mathscr{M} \models \mathrm{PA}$, and let x, $i \in \mathscr{M}$ be such that $\mathscr{M} \models L = x \wedge \neg \mathrm{Con}(\mathrm{PA}_i + L = x)$, then i is a nonstandard element of $\mathscr{M}$.

To define F we need to fix a (provably) infinitely repetitive primitive recursive coding of the elements of $V \cup \{0\}$; for example we can set: 'n codes y' if, by definition, $\exists x \leq n: n = 2^x(2y + 1)$.

DEFINITION 5.8. We define $F(0) = 0$. Suppose $F(m)$ has been defined for every $m \leq n$. Let $x = F(n)$. We define $F(n + 1)$ as follows:

1. Suppose that n codes an element y in $V \cup \{0\}$ such that xRy and $\mathrm{rank}(y, n) < \omega$. Define $F(n + 1) = y$.

2. Suppose that n codes an element y in $V \cup \{0\}$ such that $\neg xRy$ and xSy. Suppose that $\mathrm{rank}(y, n) < \mathrm{rank}(x, n)$. Note that if the rank of an element at stage n is less than ω, then it is less or equal to n. In particular, $\mathrm{rank}(y, n) \leq n$. So $a = F(\mathrm{rank}(y, n))$ has already been defined. If aRy, we define $F(n + 1) = y$.

3. In the remaining cases we define $F(n + 1) = F(n)$.

EXAMPLE 5.9. As an intuitive illustration consider the simplified ILM-model of Example 4.7 with four nodes 1, 2, 3, 4. First notice that $F: \omega \to V \cup \{0\}$ is only allowed to move along the R-arrows or the S-arrows. If xRy, then to move from x to y, F only needs to see an inconsistency from $\mathrm{PA} + L = y$. This will ensure that $\mathrm{PA} + L = x \vdash \mathrm{Con}(\mathrm{PA} + L = y)$ (as in Solovay's proof of the arithmetical completeness theorem for provability logic). On the other hand, to make an S-move, F needs to see a proof of an inconsistency from $\mathrm{PA}_i + L = y$ for a suitably small i. Consider the following situation: if F ever reaches the node 3, then it is certain that F will never assume the value 1 afterward. It will follow that $\mathrm{PA} + L = 3 \vdash \neg \mathrm{Con}(\mathrm{PA} + L = 1)$. For the nodes 3 and 4 the matter is more complicated. We have $\neg(3R4)$, so we need

to show that $PA + L = 3 \vdash \neg Con(PA + L = 4)$. The trouble is that since there is an S-arrow from 3 to 4, it is not excluded that F will move from 3 to 4 (in some model), so we have conflicting requirements. The possibility of moving from 3 to 4 is needed to ensure that every model of $PA + L = 1$ thinks that $PA + L = 3$ interprets $PA + L = 4$ (or equivalently $\forall k\ PA + L = 3 \vdash Con(PA_k + L = 4)$). The idea is to allow a move from 3 to 4 but to make it so difficult that we still have $PA + L = 3 \vdash \neg Con(PA + L = 4)$. The situation is further complicated by the fact that in any model of $PA + L = 2$ we do not want that $PA + L = 3$ interprets $PA + L = 4$ (in these models we want instead that $PA + L = 3$ is consistent and $PA + L = 4$ is not). The solution is to allow a move from 3 to 4 at stage n only if the element $a = \text{rank}(4, n)$ is so small that $F(a) \in \{0, 1\}$, namely F was still confined to the set $\{0, 1\}$ at stage a (we also need $\text{rank}(4, n) < \text{rank}(3, n)$). This 'confinement' is automatically satisfied in any model of $PA + L = 1$, and so it does not constitute a restriction; on the other hand it is a serious restriction in any model of $PA + L = 2$.

We will prove that F (and L) have the desired properties. The fact that F is a primitive recursive function is clear from the definition. It is also clear that if F has a limit, then L is the limit.

PROPOSITION 5.10 (PA). $n \leq m \to F(n)SF(m)$.

PROOF. According to which of the three clauses in the definition of F is satisfied, one of the following must hold: $F(n)RF(n + 1)$, $F(n)SF(n + 1)$, or $F(n) = F(n + 1)$. Now the thesis follows from the fact that $R \subseteq S$ and S is reflexive and transitive. QED.

PROPOSITION 5.11 (PA). *F has a limit.*

PROOF. Since the model $\mathbf{V}$ might be infinite, this is not completely clear. We know however that there is some bound k such that every R-chain has length $\leq k$. Thus F cannot make more than k consecutive R-moves. Moreover, since $xSyRz \Rightarrow xRz$, F cannot make more than k R-moves, whether they are consecutive or not. So after a certain stage, F is only allowed to make S-moves. But an S-move from a node x to a node y (at stage n) entails that the rank of y (at any stage bigger than n) is smaller than the rank of x (at stage n). So if F did not have a limit, we would have an infinite (definable) descending sequence of ranks, which is impossible. QED.

PROPOSITION 5.12 (PA). *For all x, y in $V \cup \{0\}$, if $L = x$ and xRy, then* $Con(PA + L = y)$.

PROOF. Here we use the fact that xRy implies $x \neq y$. For a contradiction, suppose that xRy, $L = x$, and $PA + L = y$ is inconsistent. Let n be such that n codes y and n is so large that: 1) F has already reached its limit x by stage n; 2) there is a proof of $L \neq y$ from PA_n with Gödel number less than n (thus $\text{rank}(y, n) < \omega$). The definition of F now entails that $F(n + 1) = y$. Absurd. QED.

PROPOSITION 5.13 (PA). *For all $x \in V$, if $L = x$, then* $\neg Con(PA + \exists y: L = y \wedge \neg xRy)$.

PROOF. Let $x = F(k)$. So $\text{rank}(x, k) \leq k$. For a contradiction, suppose that there is a model $\mathscr{Y}$ of PA and $\exists y \in \mathscr{Y}$ such that $\mathscr{Y} \models L = y \wedge \neg xRy$ (as before, the use of models is justified by working in ACA_0). Since '$x = F(k)$' is a Σ^0_1-assertion, it must hold in $\mathscr{Y}$ as well. By Proposition 5.10, $\mathscr{Y} \models xSy$. In $\mathscr{Y}$ consider the last step taken by F before reaching the limit y, namely consider n and w with $F(n) = w$, $w \neq y$,

and $\forall m > n\ F(m) = y$. Since $L = x$, n must be a nonstandard element of $\mathscr{Y}$. In particular $n > k$ (in $\mathscr{Y}$), so $\mathscr{Y} \models xSwSy$. We cannot have wRy, because this would imply xRy. Let $a = F(\text{rank}(y, n))$. By the definition of F we must have $\mathscr{Y} \models aRy$. If $\mathscr{Y} \models k \leq \text{rank}(y, n)$, by Proposition 5.10 we would have $\mathscr{Y} \models xSaRy$; hence $\mathscr{Y} \models xRy$, contradicting our assumptions. So $\mathscr{Y} \models \text{rank}(y, n) < k$, and therefore $\mathscr{Y} \models \neg\text{Con}(\text{PA}_k + L = y)$. But this is absurd by the reflection theorem, since k is a standard element of $\mathscr{Y}$. QED.

PROPOSITION 5.14 (PA). *For all $x \in V \cup \{0\}$ if $L = x$, then, for all k,* PA *proves that for all $y, z \in V \cup \{0\}$ if $L = y$, xRz, and ySz, then* $\text{Con}(\text{PA}_k + L = z)$.

PROOF. Assume that $L = x$. Fix k. For a contradiction, there is a model $\mathscr{Y}$ of PA and two elements $y, z \in \mathscr{Y}$ such that $\mathscr{Y} \models L = y \wedge xRz \wedge ySz \wedge \neg\text{Con}(\text{PA}_k + L = z)$. But then $\mathscr{Y}$ thinks that for all sufficiently large m, $\text{rank}(z, m) \leq k$. On the other hand, since k is a standard element of $\mathscr{Y}$, by the reflection theorem $\mathscr{Y}$ thinks that the theory $\text{PA}_k + L = y$ is consistent, and therefore $\forall m\ \text{rank}(y, m) > k$. Let n be such that F has already reached its limit y by stage n, n codes z, and $\text{rank}(z, n) \leq k$. In particular, $\mathscr{Y} \models \text{rank}(z, n) < \text{rank}(y, n)$. To reach a contradiction we will show that $\mathscr{Y} \models F(n + 1) = z$. Let $r \in \mathscr{Y}$ be such that $\mathscr{Y} \models r = \text{rank}(z, n)$. Note that $r \leq k$, so r must be a standard element of $\mathscr{Y}$ (although n might be nonstandard). Therefore we can compute, outside of $\mathscr{Y}$, the element $a = F(r)$. Since $L = x$, this implies that aSx (by Proposition 5.10); hence by absoluteness $\mathscr{Y} \models aSx$. Now using the property $aSxRz \rightarrow aRz$, we can conlude, in $\mathscr{Y}$, that aRz. But now clause 2 in the definition of F ensures that F will make an S-move from y to z at stage n, namely $\mathscr{Y} \models F(n + 1) = z$, which is the desired contradiction. QED.

This completes the proof of Theorem 3.8.

§6. Arithmetical completeness of ILM$^\omega$. In this section we prove that the ω-valid modal formulas are exactly the theorems of ILM$^\omega$.

DEFINITION 6.1 (Sound models). Let $\mathbf{V} = \langle V, R, S, b, \Vdash \rangle$ be a simplified ILM-model and let C be a modal formula. We say that $\mathbf{V}$ is *C-sound* if:

1. for every subformula $\Box A$ of C, $b \Vdash \Box A \rightarrow A$, and
2. for every subformula $A \rhd B$ of C, $b \Vdash \Box\neg A \rightarrow \neg A$.

Fix a provably primitive recursive ILM-model $\mathbf{V} = \langle V, R, S, b, \Vdash \rangle$. Without loss of generality we assume that $b = 1$ and $0 \notin V$, and we adjoin a new root 0 to $\mathbf{V}$ as in Definition 5.1.

PROPOSITION 6.2. *If $\mathbf{V}$ is C-sound, then, for every subformula D of C, $0 \Vdash D \Leftrightarrow 1 \Vdash D$.*

PROOF. The proof is by induction on the complexity of D. If D is atomic, the thesis is true by definition. If D is a Boolean combination of simpler formulas, the thesis follows immediately from the induction hypothesis.

Suppose that $D = A \rhd B$. It is easy to see that if $0 \Vdash D$, then $1 \Vdash D$. In fact, suppose that $0 \Vdash D$ and let y be such that $1Ry \wedge y \Vdash A$. We must show that $\exists z\colon 1Rz \wedge ySz \wedge z \Vdash B$. From the assumption $0 \Vdash D$ we can conclude that $\exists z\colon 0Rz \wedge ySz \wedge z \Vdash B$. Now since 1 is the root of V, $1RySz$ implies that z is an element of V different from 1, and therefore $1Rz$ as desired.

Conversely, suppose that $1 \Vdash D$. We need to show that $A \rhd B$ holds at 0. To see this, take an element y such that $0Ry$ (i.e. $y \in V$) and $y \Vdash A$. We must find an element

z such that $0Rz$, ySz and $z \Vdash B$. If $y \neq 1$, then we must have $1Ry$, and therefore the existence of z follows immediately from the fact that $1 \Vdash A \rhd B$ (using $1Rz \to 0Rz$). So assume $y = 1$. Since $\mathbf{V}$ is C-sound, $1 \Vdash \Box\neg A \to \neg A$. By assumption $1 \Vdash A$, so $1 \Vdash \neg\Box\neg A$. Thus A must hold at some node y' such that $1Ry'$. Since $1 \Vdash A \rhd B$, there is some z such that $1Rz \wedge y'Sz \wedge z \Vdash B$. But this implies $0Rz \wedge 1Sz$, so we are done.

The case when $D = \Box A$ can be proved similarly, using the soundness assumption $1 \Vdash \Box A \to A$. In fact if $1 \Vdash \Box A$, then by definition $\forall x(1Rx \to x \Vdash A)$. By the assumption we must also have $1 \Vdash A$. So A holds at every node x with $0Rx$, and therefore $0 \Vdash \Box A$. The converse is clear. QED.

It is clear that if $\mathbf{V}$ is provably primitive recursive the above proof can be formalized in PA. Now let I be the interpretation induced by $\mathbf{V}$ as in Definition 5.1.

DEFINITION 6.3. Let A be a modal formula. Working in PA, we say that I is *ω-faithful* on A if for all $x \in V \cup \{0\}$ we have

1. if $x \Vdash A$ and $L = x$, then A^I, and
2. if $x \Vdash \neg A$ and $L = x$, then $\neg A^I$.

LEMMA 6.4. *If $\mathbf{V}$ is C-sound, then for every subformula D of C,* PA *proves that I is ω-faithful on D.*

The proof is similar to the proof of the corresponding Lemma 5.6 with some minor modifications. Consider the crucial case $D = A \rhd B$. The only possible source of trouble in adapting the proof is that property $(\neg R)$ might no longer be applicable in the present context, since we are now dealing with all the nodes in $V \cup \{0\}$ rather than restricting our attention only to those nodes which are in V. The only places in the proof of Lemma 5.6 where we used property $(\neg R)$ were isolated in two claims:

1. There is an element $y \in \mathscr{Y}$ such that $\mathscr{Y} \models xRy \wedge y \Vdash A$.
2. Let $z \in \mathscr{Z}$ be such that $\mathscr{Z} \models L = z$. Then $\mathscr{Z} \models z \in V \wedge ySz \wedge xRz$.

In the context of Lemma 5.6 these claims were proved under the assumption that $x \in V$. In the present context we have the weaker assumption $x \in V \cup \{0\}$, but we can use the soundness property. We recall that the first claim was proved taking $y \in \mathscr{Y}$ such that $\mathscr{Y} \models L = y$ and noting that since $L = x$ and $x \in V$, by property $(\neg R)$ we must have $\mathscr{Y} \models xRy$. In the present context we can still say that $\mathscr{Y} \models (y \Vdash A)$ (using $\mathscr{Y} \models A^I$ and the ω-faithfulness on A); however if $x = 0$, nothing excludes the possibility that $y = 0$ and we cannot conclude that $\mathscr{Y} \models xRy$. But since $\mathbf{V}$ is C-sound, $0 \Vdash A$ iff $1 \Vdash A$. Thus in the unfortunate case that $L = 0$ and $\mathscr{Y} \models L = 0$ we can take $y = 1$ and we still have $\mathscr{Y} \models xRy \wedge y \Vdash A$, as desired.

To prove the second claim in the case when $x = 0$ we first notice that exactly the same argument as in the old proof shows that $\mathscr{Z} \models z \in V \wedge ySz$. But $z \in V$ implies $0Rz$, so we are done (we did not even use the soundness property).

THEOREM 6.5 (Reduction of the ω-valid formulas to ILM). *Let C be a modal formula, and let $T(C)$ be the conjunction of*: 1) *all the formulas of the form $(\Box\neg A) \to \neg A$ such that, for some B, $A \rhd B$ is a subformula of C; and* 2) *all the formulas of the form $\Box A \to A$ such that $\Box A$ is a subformula of C. Then C is ω-valid iff* $\mathrm{ILM} \vdash T(C) \to C$.

PROOF. Since all the theorems of ILM, as well as all the formulas of the form $\Box D \to D$, are ω-valid, it is clear that if $\mathrm{ILM} \vdash T(C) \to C$, then C is ω-valid. Conversely, if $\mathrm{ILM} \nvdash T(C) \to C$, then by the modal completeness theorem for ILM

there is a provably primitive recursive ILM-model $\mathbf{V}$, with root 1, such that $1 \Vdash T(C)$ and $1 \Vdash \neg C$. The fact that $1 \Vdash T(C)$ simply means that $\mathbf{V}$ is C-sound. Therefore, extending $\mathbf{V}$ by adding a new root 0 (as explained above), we have $0 \Vdash \neg C$. Since the interpretation I is ω-faithful on C, $\mathbf{PA} + L = 0 \vdash \neg C^I$. But in the standard model $L = 0$; thus $\omega \models \neg C^I$ and C is not ω-valid. QED.

We recall that $\mathbf{ILM}^\omega$ is defined as the modal theory whose axioms are all the theorems of ILM plus all the formulas of the form $\Box A \to A$, and whose sole rule of inference is modus ponens. An immediate corollary of the above theorem is that the ω-valid formulas are exactly the theorems of $\mathbf{ILM}^\omega$. This completes the proof of Theorem 3.10.

EXAMPLE 6.6. In [9] Lindström shows that there are two sentences α and β such that PA interprets both $\mathbf{PA} + \alpha$ and $\mathbf{PA} + \beta$ but it does not interpret $\mathbf{PA} + \alpha \wedge \beta$, and moreover PA does not interpret either $\mathbf{PA} + \neg\alpha$ or $\mathbf{PA} + \neg\beta$ and yet it does interpret $\mathbf{PA} + \neg\alpha \vee \neg\beta$. This can be alternatively proved using the arithmetical completeness theorem for $\mathbf{ILM}^\omega$ as follows. Let $\top$ be $\neg\bot$, and let C be the modal formula $(\top \rhd A) \wedge (\top \rhd B) \wedge \neg(\top \rhd A \wedge B) \wedge \neg(\top \rhd \neg A) \wedge \neg(\top \rhd \neg B) \wedge (\top \rhd \neg A \vee \neg B)$ (A and B are atomic formulas). Then the simplified ILM-model in Figure 2 is a C-sound model, and the formula C holds at the root.

So we can take $\alpha = A^I$ and $\beta = B^I$. Notice that in general the counterexamples provided by the induced interpretation 'I' are Σ^0_2-statements. To see this we recall that, by Definition 5.2, A^I is the formula $\exists x \in V \cup \{0\}: L = x \wedge x \Vdash A$, where $L = x$ is equivalent to the Σ^0_2-statement $\exists n \forall m \geq n\ F(m) = x$.

§7. Concluding remarks. It is easy to verify that Theorem 3.8 continues to hold if we replace the base theory PA by any sequential theory T which satisfies full induction (as explained in the Introduction) and such that T does not prove false Π^0_2-statements (this hypothesis was implicitly used in Remark 5.3). Actually it is enough to assume that T does not prove false Σ^0_1-statements, since this implies that T does not prove false Π^0_2-statements. In particular our results hold for Zermelo-Fraenkel set theory. To carry out the proof, Theorem 2.13 has to be modified as follows:

THEOREM 7.1. *If T is a sequential theory satisfying full induction, then $T + \alpha$ interprets $T + \beta$ iff for every model $\mathcal{M}$ of $T + \alpha$ there is a model $\mathcal{U}$ of $T + \beta$ such that the integers of $\mathcal{U}$ are an end-extension of the integers of $\mathcal{M}$.*

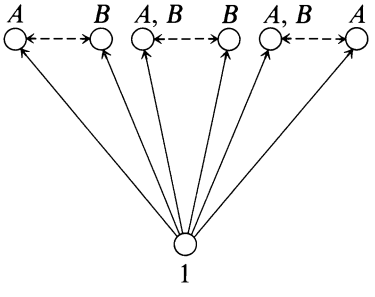

FIGURE 2

As mentioned in the Introduction, if the base theory is finitely axiomatizable we get a different interpretability logic. Let ILP be the theory obtained from ILM by replacing Montagna's principle by the axiom scheme $A \rhd B \to \Box(A \rhd B)$. Visser and Smoryński [18] proved that ILP is the interpretability logic of the finitely axiomatizable theories GB and ACA_0. (Namely, $\mathrm{ILP} \vdash A$ iff for every interpretation f based on GB, $\mathrm{GB} \vdash A^f$; similarly for ACA_0.) To explain the soundness of the new axiom it is enough to observe that for a fixed finitely axiomatizable theory T the complexity of the assertion '$T + \alpha$ interprets $T + \beta$' is Σ^0_1 (as α and β range over sentences of T); therefore its truth implies its provability. (For $T = \mathrm{PA}$ the corresponding assertion is Π^0_2 by Orey's theorem.)

Let SUPEXP be a formalization of the assertion that superexponentiation is a total function. (On input n this function yields a stack of n two's with an n on top. This rate of growth is needed to prove the cut elimination theorem.) In [19] Visser generalized the completeness theorem for GB and ACA_0 to any finitely axiomatizable sequential theory T containing $\mathrm{I}\Delta_0 + \mathrm{SUPEXP}$ which does not prove false Σ^0_1-sentences (actually it suffices that T does not prove its own 'iterated inconsistency' for any number of iterations).

It is still open what is the interpretability logic of weaker theories: De Jongh and Veltman found an example showing that the interpretability logic of $\mathrm{I}\Delta_0 + \mathrm{EXP}$ properly extends ILP.

Visser's result shows in particular that the language of interpretability logic is still too weak to distinguish between, say, $\mathrm{I}\Sigma_1$ and GB, despite the fact that the latter theory is a predicative extension of an essentially reflexive theory (ZF) and the former is not.

This and other considerations (like Rosser's sentences) raise the issue of expanding the modal language to obtain stronger results.

An obvious try in this direction would be the introduction of quantifiers over numbers (as in $\Box\forall x A(x) \to \forall x \Box A(x)$). However Vardanyan [17] proved that this leads to an undecidable set of PA-valid formulas (even without the $\rhd$-operator), and Artemov [1] showed that the corresponding set of ω-valid formulas is not even arithmetical. In [3] we proved that Vardanyan's undecidability result still holds even if we restrict the interpretation of the atomic modal formulas to belong to the set Σ^0_1. (Although we get new provability principles, like $A \to \Box A$.)

A different approach would be that of expanding the language with the introduction of quantifiers over proposition. We do not know whether this leads to an undecidable theory; if it does, probably some restricted version of this proposal is more feasible: for example one could introduce a new modal operator '$\sqsupset$', where the intended meaning of $A \sqsupset B$ is 'there is a (finite) consistent extension of $\mathrm{PA} + A$ which interprets $\mathrm{PA} + B$'. (In a private communication Solovay observed that if there is such an extension, then there is a finite one. This also follows from some results of Lindström [8].)

A third possibility that has already proved useful in the study of the modal logic of provability (see for example Carlson's completeness theorem in [15, p. 206]) is to have simultaneously in the modal language different provability and interpretability predicates corresponding to different base theories.

For other open problems we refer the reader to [18].

Appendix A. The completeness theorem of de Jongh and Veltman. To keep this paper self-contained we give a proof of the theorem of de Jongh and Veltman [5] which establishes the completeness of ILM with respect to a class of finite models (and therefore the decidability). The results of this section will be used in Appendix B. We follow very closely the exposition of De Jongh and Veltman except for the fact that we use a slight variation of their definition of 'adequate set' suggested by Solovay.

DEFINITION A.1. An ILM-*frame* is an L-frame $\langle W, R, b\rangle$ together with an additional relation S_w for each $w \in W$, with the following properties:

1. $uS_wv \Rightarrow wRu \wedge wRv$;
2. S_w is reflexive and transitive;
3. if $wRuRv$, then uS_wv; and
4. if uS_wvRz, then uRz.

DEFINITION A.2. An ILM-*model* is an ILM-frame together with a forcing relation '$\Vdash$' which satisfies the usual clauses for the propositional connectives and the $\Box$-operator plus the following:

$$u \Vdash A \rhd B \Leftrightarrow \forall v(uRv \wedge v \Vdash A \Rightarrow \exists w(vS_uw \wedge w \Vdash B)).$$

It is easy to verify that ILM is sound with respect to ILM-models; that is, if $\mathrm{ILM} \vdash A$, then, for every ILM-model $\mathbf{W}$, $\mathbf{W} \models A$. Dick de Jongh and Frank Veltman [5] prove that, conversely, if $\mathrm{ILM} \nvdash A$, then there is a finite ILM-model $\mathbf{W}$, with root b say, such that $b \Vdash \neg A$. So the theory ILM is complete with respect to the class of all the finite ILM-models, and therefore it is decidable. The remainder of this Appendix is devoted to giving a proof of this fact. The elements of the ILM-models that we are going to construct are ordered pairs $\langle \Delta, \tau\rangle$ whose first component is a finite consistent set of modal formulas satisfying certain closure properties. We need some technical lemmas to prove that such sets exist.

DEFINITION A.3. An *adequate* set of formulas is a set Φ which fulfills the following conditions:

1. Φ is closed under the taking of subformulas.
2. If $B \in \Phi$ and B is not a negated formula, then $\neg B \in \Phi$.
3. $\bot \rhd \bot \in \Phi$.
4. If $B \rhd C \in \Phi$, then also $\Diamond B, \Diamond C \in \Phi$.
5. If B as well as C is an antecedent or a consequent of some $\rhd$-formula in Φ, then $B \rhd C \in \Phi$.
6. If $B \rhd C, \Box D \in \Phi$, then there are formulas B' and C', which are L-equivalent to $B \wedge \Box D$ and $C \wedge \Box D$ respectively, such that $B' \rhd C' \in \Phi$.

We will show that every finite set of formulas is contained in a finite adequate set.

LEMMA A.4. *Let U be a finite set of formulas. Consider the operation $f(A, B) = A \wedge \Box\neg B$. Let X be the smallest set of formulas containing U and closed under f. Clearly X is infinite. We claim that X is included in only finitely many equivalence classes with respect to L-provable equivalence.*

PROOF. The lemma clearly follows from the special case in which $\bot \in U$ and all the other formulas of U are propositional variables. The proof of this special case proceeds by induction on the cardinality of U. Note that a formula C belongs to X if and only if C has the form $C = A \wedge \Box\neg D_1 \wedge \cdots \wedge \Box\neg D_n$, where $A \in U$ and

all the D's are in X (possibly $n = 0$). Therefore it is enough to show that up to L-equivalence there are only finitely many formulas of the form $\Box\neg D$ with $D \in X$ (granted this, we have a bound on n as well). If $U = \{\bot\}$, then every formula in X is L-equivalent to $\bot$ and we are done. So assume $|U| > 1$. Consider $D \in X$. Then there is $n \geq 0$ such that D has the form $u \wedge E$, where $E = \Box\neg F_1 \wedge \cdots \wedge \Box\neg F_n$ and all the F's are in X (we can disregard the parentheses, since the conjunction is associative). If $u = \bot$, then $\Box\neg D$ is L-equivalent to $\Box\neg\bot$. If $u \neq \bot$, let $E' = \Box\neg F'_1 \wedge \cdots \wedge \Box\neg F'_n$ be obtained from E by replacing all the occurrences of u in E by $\bot$. By the induction hypothesis there are only finitely many possible choices for E' up to L-equivalence (warning: E' does not belong to X, but we can apply the induction hypothesis to its subformulas F'_i). Therefore to prove the lemma it is enough to show that $\Box\neg(u \wedge E)$ is L-equivalent to $\Box\neg(u \wedge E')$. To see this we use the completeness Theorem 4.3 of L with respect to finite Kripke models. First note that E, being a conjunction of $\Box$-formulas, is preserved upwards in a Kripke model; that is, if $x \Vdash E$ and xRy, then $y \Vdash E$. Now suppose that $\Box\neg(u \wedge E)$ fails in a (finite) Kripke model for L with root b. This means that there is a node $x \neq b$ such that $x \Vdash u \wedge E$. We can assume that x is an R-maximal such node. So we must have $x \Vdash u \wedge \Box\neg u \wedge E$. But this implies $x \Vdash u \wedge E'$ (since every occurrence of u in E lies within the scope of a $\Box$-operator); hence $\Box\neg(u \wedge E')$ fails in the Kripke model as well. The converse is completely similar. QED.

LEMMA A.5. *Every finite set of formulas is contained in a finite adequate set.*

PROOF. Let Φ_1 be a finite set of formulas. We can assume that $\bot \rhd \bot \in \Phi_1$. Define '$\sim A$' as follows: if A is not a negated formula then $\sim A$ is $\neg A$; if $A = \neg B$, $\sim A$ is B. We call $\sim A$ the *pseudonegation* of A. Let U_0 be the closure of Φ_1 under subformulas, and let U be the closure of U_0 under pseudonegations. Then U is closed under both subformulas and pseudonegations. Let X be the union of an infinite sequence of sets $X_0, X_1, X_2, \ldots$, where X_0 is U and X_{n+1} is the union of X_n and the set of all the formulas of the form $F \wedge \Box\neg G$, with $F, G \in X_n$, which are not L-equivalent to any formula in X_n. Clearly X is closed under the function $f(G, B) = G \wedge \Box\neg B$ up to provable equivalence in L. By the previous lemma X is finite. Now let

$$\Phi_2 = U \cup \{B \rhd C \mid B, C \in X\} \cup \{\Box\neg A \mid A \in X\}.$$

Finally, let Φ_3 be the closure of Φ_2 under subformulas and pseudonegations. We claim that Φ_3 is a finite adequate set containing Φ_1.

To prove this we need the following facts that can be easily proved by induction on n:

1. If a formula of the form $B \rhd C$ is a subformula of a formula in X_n, then $B \rhd C \in U$.

2. If a formula of the form $\Box D$ is a subformula of a formula in X_{n+1}, then either $D = \neg A$ for some $A \in X_n$, or $\Box D \in U$.

Now to prove clause 6 of the definition of 'adequate set' suppose that $B \rhd C$ and $\Box D$ are in Φ_3. Then $B \rhd C$ is a subformula of a formula in Φ_2. So by the definition of Φ_2 one of the following holds: 1) $B \rhd C$ is a subformula of a formula in U, or 2) $B, C \in X$, or 3) $B \rhd C$ is a subformula of a formula in X and therefore belongs to U. Since $X \supseteq U$ and U is closed under subformulas, in any case $B, C \in X$.

Similarly from the fact that $\Box D \in \Phi_3$ it follows that $\Box D$ is a subformula of a formula in Φ_2. By the definition of Φ_2 and a previous remark we can conclude that either $D = \neg A$ with $A \in X$, or $\Box D \in U$. Since U is closed under subformulas and pseudonegations, in either case $\neg D$ is equivalent to some formula in X. Since X is closed under f up to L-equivalence, the formulas $B \wedge \Box D$ and $C \wedge \Box D$ are L-equivalent to some formulas in X (hence in Φ_3). This proves clause 6.

To prove clause 5, note that if B as well as C is an antecedent or a consequent of some $\triangleright$-formula in Φ_3, then, reasoning as above, B, $C \in X$, and therefore $B \triangleright C \in \Phi_2$. But $\Phi_2 \subseteq \Phi_3$, so we are done.

For clause 4 we recall that '$\Diamond$' is defined as '$\neg\Box\neg$'; now if $B \triangleright C \in \Phi_3$, then B, $C \in X$, hence $\Box\neg B$ and $\Box\neg C$ are in Φ_2, and therefore their negations are in Φ_3 as desired. The other clauses are easy to verify. QED.

Given a set of modal formulas Γ, we write $\Gamma \vdash A$ if there is a finite conjunction C of formulas from Γ such that $\mathrm{ILM} \vdash C \to A$. We say that Γ is ILM-*consistent* if $\Gamma \nvdash \bot$. In the following discussion we consider a fixed finite adequate set Φ and we let 'Γ' and 'Δ' denote maximal ILM-consistent subsets of Φ.

DEFINITION A.6. We write $\Gamma \prec \Delta$ (Δ is a successor of Γ) iff:

1. A, $\Box A \in \Delta$ for each $\Box A \in \Gamma$, and
2. $\Box A \in \Delta$ for some $\Box A \notin \Gamma$.

DEFINITION A.7. Let C be a formula. We say that Δ is a C-critical successor of Γ iff:

1. $\Gamma \prec \Delta$, and
2. $\neg A$, $\Box\neg A \in \Delta$ for each A such that $A \triangleright C \in \Gamma$ (Δ contains no formula that 'asks' for C).

Note that a successor of a C-critical successor is a C-critical successor.

LEMMA A.8. *The following are theorems of* ILM:

1. $(\Box\neg B) \to (B \triangleright C)$.
2. $A \triangleright (A \wedge \Box\neg A)$.

PROOF. First we prove 2. The formula $\Box(\Diamond A \to \Diamond(A \wedge \Box\neg A))$ is a theorem of the modal logic of provability (which is included in ILM). Now reason in ILM (see Definition 3.7). By axiom 4 (applied to the formula above), $\Diamond A \triangleright \Diamond(A \wedge \Box\neg A)$. Now by the transitivity of '$\triangleright$' and axiom 8, we have $\Diamond A \triangleright (A \wedge \Box\neg A)$, namely $\neg\Box\neg A \triangleright (A \wedge \Box\neg A)$. Using axiom 6, we can combine this with the fact that $(A \wedge \Box\neg A) \triangleright (A \wedge \Box\neg A)$, to get the desired conclusion: $A \triangleright (A \wedge \Box\neg A)$.

To prove 1, first note that $\Box\neg B \to \Box(B \to C)$ is a theorem of the logic of provability. Now by axiom 4 of ILM (and tautological reasoning), ILM proves $\Box\neg B \to (B \triangleright C)$. QED.

LEMMA A.9. *If* $\neg(B \triangleright C) \in \Gamma$, *then* Γ *has a* C*-critical successor* Δ *such that* $B \in \Delta$.

PROOF. Since Γ is ILM-consistent, by this lemma $\Box\neg B \notin \Gamma$. Let Ψ be the set

$$\{D, \Box D \mid \Box D \in \Gamma\} \cup \{\neg A, \Box\neg A \mid A \triangleright C \in \Gamma\} \cup \{B, \Box\neg B\}.$$

By the adequacy conditions Ψ is included in Φ. We will show that Ψ is ILM-consistent. Granted this, we can take Δ to be a Φ-completion of Ψ (namely a maximal ILM-consistent subset of Φ containing Ψ) and we are done. Suppose, for a contradiction, that Ψ is not ILM-consistent. Then we can write

$$D_1, \ldots, D_k, \Box D_1, \ldots, \Box D_k \vdash B \wedge \Box\neg B \to A_1 \vee \cdots \vee A_m \vee \Diamond(A_1 \vee \cdots \vee A_m),$$

where $\Box D_i \in \Gamma$ and $A_i \rhd C \in \Gamma$ (we agree that the empty disjunction is $\bot$ and the empty conjunction is $\neg\bot$). Now, since ILM contains L,

$$\Box D_1, \ldots, \Box D_k \vdash \Box(B \wedge \Box\neg B \to A_1 \vee \cdots \vee A_m \vee \Diamond(A_1 \vee \cdots \vee A_m)).$$

By the axiom $\Box(U \to V) \to U \rhd V$,

$$\Box D_1, \ldots, \Box D_k \vdash (B \wedge \Box\neg B) \rhd (A_1 \vee \cdots \vee A_m \vee \Diamond(A_1 \vee \cdots \vee A_m)).$$

Since $U \vee \Diamond U \rhd U$ is a theorem of ILM,

$$\Box D_1, \ldots, \Box D_k \vdash B \wedge \Box\neg B \rhd A_1 \vee \cdots \vee A_m.$$

Now, using the fact that $\Box D_i \in \Gamma$ and $B \rhd (B \wedge \Box\neg B)$ is a theorem of ILM,

$$\Gamma \vdash B \rhd A_1 \vee \cdots \vee A_m.$$

Finally, since, for each i, $A_i \rhd C \in \Gamma$,

$$\Gamma \vdash B \rhd C.$$

This contradicts the consistency of Γ. QED.

LEMMA A.10. *Suppose $B \rhd C \in \Gamma$ and let Δ be an E-critical successor of Γ with $B \in \Delta$. Then there is an E-critical successor Δ' of Γ with $C \in \Delta'$.*

PROOF. Let Ψ be the set

$$\{D, \Box D \mid \Box D \in \Gamma\} \cup \{\neg F, \Box\neg F \mid F \rhd E \in \Gamma\} \cup \{C, \Box\neg C\}.$$

By the adequacy conditions $\Psi \subseteq \Phi$. If $\Box\neg C \in \Gamma$, then $\Box\neg B \in \Gamma$ (as $B \rhd C \in \Gamma$), contradicting $B \in \Delta$. So $\Box\neg C \notin \Gamma$. Suppose that Ψ is inconsistent. Then, reasoning as above, $\Gamma \vdash C \rhd F_1 \vee \cdots \vee F_m$ (where, for each i, $F_i \rhd E \in \Gamma$); hence $\Gamma \vdash B \rhd F_1 \vee \cdots \vee F_m$. If $m = 0$ we would have $\Gamma \vdash \Box\neg B$, contradicting the fact that Δ is a (consistent) successor of Γ with $B \in \Delta$. So $m > 0$. This means in particular that E is the consequent of some formula in Φ. Since, for each i, $F_i \rhd E \in \Gamma$, we have $\Gamma \vdash B \rhd E$. Since $B \rhd C \in \Gamma$, B is the antecedent of a $\rhd$-formula in Φ; so, by the adequacy conditions, $B \rhd E \in \Phi$. Hence $\Gamma \vdash B \rhd E$ can be strengthened to $B \rhd E \in \Gamma$. Since Δ is an E-critical successor of Γ, this implies $\neg B \in \Delta$, contradicting our assumptions. Thus we have proved that Ψ is consistent. Let Δ' be a Φ-completion of Ψ. Then Δ' has the desired properties. QED.

THEOREM A.11. *If* ILM $\nvdash A$, *then there is a finite* ILM*-model* $\mathbf{W}$, *with root b say, such that* $b \Vdash \neg A$.

PROOF. Take some finite ILM-adequate set Φ containing $\neg A$ and let Γ be a maximal ILM-consistent subset of Φ containing $\neg A$. The *depth* of Γ is by definition the length of the longest chain $\Gamma = \Gamma_1 \prec \cdots \prec \Gamma_n$. The underlying set of the model $\mathbf{W}$ is defined as the set W_Γ consisting of all the pairs $\langle \Delta, \tau \rangle$ such that

1. Δ is a maximal ILM-consistent subset of Φ such that $\Gamma \prec \Delta$ or $\Gamma = \Delta$, and
2. τ is a finite sequence of formulas from Φ, the length of which does not exceed the depth of Γ minus the depth of Δ. (So $\langle \Gamma, \tau \rangle \in W_\Gamma$ iff τ is the empty sequence.)

Given a pair $w = \langle \Delta, \tau \rangle$, we denote by $(w)_0$ and $(w)_1$ the first and the second component of w respectively. Define R on W_Γ as follows: $wRw' \Leftrightarrow (w)_0 \prec (w')_0 \wedge (w)_1 \subseteq (w')_1$. We say that u is a *C-critical R-successor* of w if $(u)_0$ is a C-critical successor of $(w)_0$ and $(u)_1$ has the form $(w)_1 * \langle C \rangle * \tau$ (so C is uniquely determined

by u). Define $uS_w v$ if the following hold:

1. wRu and wRv.
2. $(u)_1 \subseteq (v)_1$.
3. For each A such that $\Box A \in (u)_0$, also $\Box A \in (v)_0$.
4. If u is a C-critical R-successor of w, so is v.

For p atomic, define $w \Vdash p$ iff $p \in (w)_0$. It is easy to verify that W_Γ, equipped with the accessibility relations R and S_w, and with the forcing relation $\Vdash$, is a finite ILM-model with root $b = \langle \Gamma, \varnothing \rangle$. To finish the proof of the completeness theorem it is enough to show that, for each $A \in \Phi$, $w \Vdash A \Leftrightarrow A \in (w)_0$. The proof is by induction on the complexity of A. We restrict ourselves to the case that A is $B \rhd C$. We can do this since the equivalence $\Box A \leftrightarrow (\neg A \rhd \bot)$ is provable in ILM (and therefore also holds in our semantical interpretations). So we have to show that

$$B \rhd C \in (w)_0 \Leftrightarrow \forall u(wRu \wedge B \in (u)_0 \Rightarrow \exists v(uS_w v \wedge C \in (v)_0)).$$

$\Leftarrow$. Suppose $B \rhd C \notin (w)_0$. Then $\neg(B \rhd C) \in (w)_0$. We must show that

$$\exists u(wRu \wedge B \in (u)_0 \wedge \forall v(uS_w v \Rightarrow \neg C \in (v)_0)).$$

By Lemma A.9, $(w)_0$ has a C-critical successor u_0 with $B \in u_0$. Let $u_1 = (w)_1 * \langle C \rangle$ (the concatenation of $(w)_1$ and $\langle C \rangle$), and let $u = \langle u_0, u_1 \rangle$. Then $u \in W_\Gamma$ and u is a C-critical R-successor of w. Consider any v such that $uS_w v$. Since u is a C-critical R-successor of w, v will be one too. Therefore $\neg C \in (v)_0$, as desired.

$\Rightarrow$. Suppose $B \rhd C \in (w)_0$, and let u be such that wRu and $B \in (u)_0$. Let $\{\Box D_1, \ldots, \Box D_n\}$ be the set $\{\Box D \mid \Box D \in (u)_0\}$. Since (w_0) is a maximal ILM-consistent subset of Φ, Montagna's principle (cf. §3) and the adequacy of Φ insure that there are two formulas B', $C' \in \Phi$, which are L-equivalent to $B \wedge \Box D_1 \wedge \cdots \wedge \Box D_n$ and $C \wedge \Box D_1 \wedge \cdots \wedge \Box D_n$ respectively, such that $B' \rhd C' \in (w)_0$. Since $B, \Box D_1, \ldots, \Box D_n \in (u)_0$, we must have $B' \in (u)_0$.

Let us first assume that, for some E, u is an E-critical R-successor of w. Then $(u)_1$ has the form $(w)_1 * \langle E \rangle * \tau$. Now by Lemma A.10 there is an E-critical successor v_0 of $(w)_0$ with $C' \in v_0$. Thus $C, \Box D_1, \ldots, \Box D_n \in v_0$. Let $v_1 = (u)_1$ and $v = \langle v_0, v_1 \rangle$. Since each $\Box$-formula in $(u)_0$ is also an element of v_0, the depth of v_0 cannot be larger than the depth of $(u)_0$. Therefore $v \in W_\Gamma$. It is easy to verify that $uS_w v$, so we are done.

If on the other hand u is not an E-critical R-successor of w, then we only know that $(w)_0 \prec (u)_0$. But every successor is a $\bot$-critical successor. So we can still apply Lemma A.10 to obtain a successor v_0 of $(w)_0$ with $C' \in v_0$. Now take $v_1 = (u)_1$ and $v = \langle v_0, v_1 \rangle$. QED.

Appendix B. Visser's simplified models. ILM-models are defined in Appendix A. Simplified ILM-models are defined in §4.2.

We will show, following Visser [18], that ILM-models and simplified ILM-models can simulate each other.

One direction is easy: it is always possible to transform a simplified ILM-model $\mathbf{V} = \langle V, R, S, b, \Vdash \rangle$ into an ILM-model with the same underlying set and the same R-relation by defining $uS_w v \Leftrightarrow wRu \wedge wRv \wedge uSv$. The forcing relation of the resulting ILM-model is defined so that it agrees on atomic formulas with the forcing relation on the simplified ILM-model $\mathbf{V}$. It is easy to verify that the forcing relations on the two models will then agree on every modal formula. So we get a

simplified ILM-model $\langle V, R, \{S_x \mid x \in V\}, b, \Vdash\rangle$. If we identify a simplified ILM-model with the induced ILM-model, we can therefore think of the simplified ILM-models as a subset of the ILM-models.

Conversely we will show that for every ILM-model there is an equivalent simplified ILM-model (in the sense that the same modal formulas will hold at the root of the two models); the two corresponding models however will not have in general the same underlying set. To explain the exact relation between the two models we have to define the notion of 'bisimulation'.

DEFINITION B.1. Consider two ILM-models $\mathbf{W}$ and $\mathbf{W}'$, and let β be a relation between W and W' (that is, $\beta \subseteq W \times W'$). We say that β is a *bisimulation* between $\mathbf{W}$ and $\mathbf{W}'$ if the following hold.

1. $b\beta b'$ (where b and b' are the roots of $\mathbf{W}$ and $\mathbf{W}'$ respectively).
2. Let x, y, z range over W and x', y', z' range over W'. For every x, x' with $x\beta x'$ we have

$$\forall y(xRy \Rightarrow \exists y'(y\beta y' \wedge x'R'y' \wedge \forall z'(y'S'_{x'}z' \rightarrow \exists z(z\beta z' \wedge yS_xz)))).$$

3. Conversely,

$$\forall y'(x'R'y' \Rightarrow \exists y(y\beta y' \wedge xRy \wedge \forall z(yS_xz \rightarrow \exists z'(z\beta z' \wedge y'S'_{x'}z')))).$$

4. $x\beta x' \Rightarrow (x \Vdash p \Leftrightarrow x' \Vdash' p)$, for all atoms p.

We say that two ILM-models *bisimulate* if there is a bisimulation between them. Clearly bisimulating is an equivalence relation between ILM-models; moreover, we have

PROPOSITION B.2. *Suppose that β is a bisimulation between $\mathbf{W}$ and $\mathbf{W}'$, and $x\beta x'$. Then, for any $A \in \mathscr{L}(\rhd)$, $x \Vdash A \Leftrightarrow x' \Vdash' A$.*

PROOF. We proceed by induction on A. We prove the case when A is $B \rhd C$. By symmetry we can assume $x \nVdash A$. So there is a y with xRy, $y \Vdash B$, and, for all z, $yS_xz \Rightarrow z \nVdash C$. Since $x\beta x'$, there is y' such that

$$y\beta y' \wedge x'R'y' \wedge \forall z'(y'S'_{x'}z' \Rightarrow \exists z(z\beta z' \wedge yS_xz)).$$

Since $y\beta y'$, by the induction hypothesis $y' \Vdash' B$. If there were a z' with $y'S'_{x'}z'$ and $z' \Vdash' C$, there would be a z with yS_xz and $z\beta z'$, and hence by the induction hypothesis $z \Vdash C$, contrary to the choice of y. Thus z' does not exist, and therefore y' witnesses the fact that $x' \nVdash' A$, as desired. QED.

THEOREM B.3. *Every finite* ILM*-model $\mathbf{W} = \langle W, R, \{S_x \mid x \in W\}, b, \Vdash\rangle$ can be bisimulated by (the* ILM*-model induced by) a simplified* ILM*-model $\mathbf{W}'$.*

It follows that we have a completeness theorem for ILM with respect to simplified ILM-models. To define $\mathbf{W}'$ we need to define its set of worlds W', a transitive conversely well-founded relation R' (on W'), a transitive reflexive relation $S' \supseteq R'$, a root b', and a forcing relation $\Vdash'$. The relation S' will be defined as the relation of end-extension among finite sequences of elements from W and therefore will have the additional property of being antisymmetric (which we do not need). We now give the definition:

DEFINITION B.4. 1. W' is the set of all those finite sequences $\langle x_1, \ldots, x_n\rangle$ of elements from W such that $x_1 = b$ and for every $i < n$ either x_iRx_{i+1} or $\exists j < i$: $x_iS_{x_j}x_{i+1}$.

2. $b' = \langle b \rangle$.

3. S' is the relation of end-extension; that is, $\langle x_1, \ldots, x_n \rangle S' \langle x_1, \ldots, y_m \rangle$ iff $m \geq n$ and, for all $i \leq n$, $x_i = y_i$.

4. $\langle x_1, \ldots, x_n \rangle R' \langle y_1, \ldots, y_m \rangle$ iff $\vec{y}$ is a proper end-extension of $\vec{x}$ and $\exists k$ with $n \leq k < m$ such that $y_k R y_{k+1}$, and for all j with $k < j < m$ there is an s with $n \leq s < j$ satisfying $y_j S_{y_s} y_{j+1}$. By considering the maximal such k we can assume without loss of generality that $\forall j > k \neg(y_j R y_{j+1})$.

5. For p atomic, $\langle x_1, \ldots, x_n \rangle \Vdash' p$ iff $x_n \Vdash p$.

It is easy to verify that $\mathbf{W}'$ is a simplified ILM-model. In particular the fact that R' is reverse well-founded follows from the reverse well-foundedness of R and the next lemma. $\mathbf{W}'$ becomes an ILM-model after defining $yS'_x z \Leftrightarrow xR'y \wedge xR'z \wedge yS'z$.

LEMMA B.5. *If* $\langle x_1, \ldots, x_n \rangle R' \langle y_1, \ldots, y_m \rangle$, *then* $x_n R y_m$.

PROOF. By the definition of R' we must have $m > n$. We have either $y_{m-1} R y_m$ or $\exists s$ with $n \leq s < m$ such that $y_{m-1} S_{y_s} y_m$, and hence $y_s R y_m$. So in either case there is some j with $n \leq j < m$ such that $y_j R y_m$. By the property $uS_w vRz \Rightarrow uRz$, we can conclude that $x_n R y_m$. QED.

LEMMA B.6. *If* $\langle \cdots x \rangle R' \langle \cdots x, y \rangle S' \langle \cdots x, y \cdots z \rangle$ *and* $\langle \cdots x \rangle R' \langle \cdots x, y \cdots z \rangle$, *then* $yS_x z$ (*possibly* $y = z$).

PROOF. Since $\langle \cdots x \rangle R' \langle \cdots x, y \rangle$, we have xRy. There are two cases to consider:

Case 1. The R-step witnessing $\langle \cdots x \rangle R' \langle \cdots x, y \cdots z \rangle$ happens between x and y. If only S_x-steps need to be taken between y and z, we are done. If not, there are u, v, w with $vS_u w$ such that $\langle \cdots z \rangle = \langle \cdots x, y \cdots u \cdots v, w \cdots z \rangle$ (possibly $y = u$, possibly $w = z$), and the last non-S_x step is taken between v and w. Since $vS_u w$, we have in particular uRw. We can go from y to u with a series of S-steps (with various subindices), and from u to w with an R-step; thus we must have yRw. But now $xRyRw$ entails $yS_x w$. From w to z only S_x-steps are taken; so, by transitivity and reflexivity of S_x, $yS_x z$.

Case 2. The R-step witnessing $\langle \cdots x \rangle R' \langle \cdots x, y \cdots z \rangle$ happens between y and z. So there are c and d such that cRd, $\langle \cdots z \rangle = \langle \cdots x, y \cdots c, d \cdots z \rangle$ (possibly $y = c$, possibly $d = z$), and from d to z there are only S_i-steps for i occurring in $x, y \cdots c, d \cdots z$. Reasoning as above, cRd implies yRd and therefore $yS_x d$. If between d and z only S_x-steps occur, we are done. Otherwise there are u, v, w, with $vS_u w$, such that $\langle \cdots z \rangle = \langle \cdots x, y \cdots u \cdots v, w \cdots z \rangle$ (possibly $y = u$, possibly $w = z$), where v occurs in $d \cdots z$ and the last non-S_x step is taken between v and w. We have uRw, so, reasoning as in Case 1, yRw holds; hence $yS_x w$. From w to z only S_x-steps are taken; so we are done. QED.

LEMMA B.7. *Define* $\beta \subseteq W \times W'$ *by* $x_n \beta \langle x_1, \ldots, x_n \rangle$. *Then* β *is a bisimulation between* $\mathbf{W}$ *and* (*the* ILM-*model induced by*) $\mathbf{W}'$.

PROOF. Clauses 1 and 4 in the definition of bisimulation are true by definition. We prove clause 2. Suppose $x\beta\langle \cdots x \rangle$ and xRy. Then $\langle \cdots x \rangle R' \langle \cdots x, y \rangle$. Suppose now that $\langle \cdots x, y \rangle S' \langle \cdots x, y \cdots z \rangle$ and $\langle \cdots x \rangle R' \langle \cdots x, y \cdots z \rangle$. It is enough to show that $yS_x z$. This is guaranteed by the previous lemma.

Finally we prove clause 3. Suppose $x\beta\langle \cdots x \rangle$ and $\langle \cdots x \rangle R' \langle \cdots x, \cdots y \rangle$. By Lemma B.5, xRy. Suppose $yS_x z$. Then $\langle \cdots x, \cdots y, z \rangle \in W'$. It is enough to show that

$\langle \cdots x\rangle R'\langle \cdots x, \cdots y, z\rangle$ and $\langle \cdots x, \cdots y\rangle S'\langle \cdots x, \cdots y, z\rangle$. The last assertion is clear. The first follows immediately from $\langle \cdots x\rangle R'\langle \cdots x, \cdots y\rangle$ and $yS_x z$. QED.

We have thus proved Theorem B.3. We will show that the model $\mathbf{W}'$ is provably primitive recursive (see Definition 4.10). It is clear that the construction of the bisimulating simplified ILM-model $\mathbf{W}'$ can be carried out in PA and that the relations S' and R' are primitive recursive. It is also clear that the forcing relation '$\Vdash'$' is primitive recursive, since it can be reduced to the forcing relation on a finite ILM-model: $\langle x_1, \ldots, x_n\rangle \Vdash' A \Leftrightarrow x_n \Vdash A$ (this follows from the fact that the two models bisimulate). Therefore it only remains to prove that the formalization of the relation '$\exists \vec{z}(\vec{x}R'\vec{z} \wedge \vec{y}S'\vec{z} \wedge \vec{z} \Vdash' B)$' is equivalent in PA to a primitive recursive formula. This will follow from the fact that we can give a bound on the length of the witness $\vec{z}$ as follows:

PROPOSITION B.8. *Let* $P(\vec{x}, \vec{y}, \vec{z})$ *be the relation* $\vec{x}R'\vec{z} \wedge \vec{y}S'\vec{z} \wedge \vec{z} \Vdash' B$. *Then there is a number* k (*independent of* B, $\vec{x}$ *and* $\vec{y}$) *such that if* $\exists \vec{z}\colon P(\vec{x}, \vec{y}, \vec{z})$, *then* $\exists \vec{z}\colon P(\vec{x}, \vec{y}, \vec{z}) \wedge |\vec{z}| \leq |\vec{y}| + k$.

PROOF. Suppose that $\vec{z}$ is a sequence of minimal length such that $P(\vec{x}, \vec{y}, \vec{z})$. Let $t = |W| + 1$. By the pigeonhole principle, in any sequence of elements from W of length t there is a repetition. $\vec{z}$ is an end-extension of $\vec{y}$, so we can write $\vec{z} = \vec{y} * \vec{u}$. Now we write $\vec{u}$ as a concatenation of sequences of length t (except possibly the last one, which might have length $< t$). Since W is finite, we can find a number k such that if $|\vec{u}| > k$, then there will be one sequence $\vec{s}$ which occurs at least 3 times in this decomposition of $\vec{u}$. We claim that the second occurrence of $\vec{s}$ can be replaced by a shorter sequence, thus contradicting the minimality of $\vec{z}$. To see this we first notice that in $\vec{s}$ there must be a repetition; namely, there must be two occurrences of the same element $a \in W$. To obtain a shorter sequence we can now delete (in the second occurrence of $\vec{s}$) one of the two occurrences of a together with all the elements of $\vec{s}$ that occur between these two occurrences of a. The fact that this process does not lead outside of the underlying set W' is guaranteed by the fact that for every node that we deleted there is an earlier occurrence of the same node (the one occurring in the first occurrence of $\vec{s}$) that we did not delete. We need to verify that the R-step witnessing the fact that $\vec{y}R'\vec{z}$ is not destroyed after the shortening of $\vec{z}$. In fact if this R-step happened in the portion of $\vec{z}$ that we deleted, namely in the second occurrence of $\vec{s}$, then the corresponding step which happens in the third occurrence of $\vec{s}$ will work as well as a witness. QED.

This completes the proof of Theorem 4.11.

Acknowledgements. The results of this paper were obtained while I was a Ph.D. student at the University of California at Berkeley. During that period I received financial support from the University of California and from the Italian National Research Institute (CNR). This paper was partially supported by NSF Grant DMS 8701828. I want to express my gratitude to my thesis advisor, Professor Robert Solovay, for the guidance received in my research and for his proofreading a preliminary version of this paper. Correspondence with Professor Albert Visser was very stimulating; in particular I want to thank him for keeping me up with his own research on the modal logic of interpretability of Peano arithmetic, and for his comments on a preliminary version of this paper. (He also suggested a simplification of a previous version of the proof of Lemma 5.13.)

Discussions with Professor Petr Hájek helped the author to improve some sections of this paper.

REFERENCES

[1] S. N. ARTEMOV, *Nonarithmeticity of truth predicate logics of provability*, ***Doklady Akademii Nauk SSSR***, vol. 284 (1985), pp. 270–271; English translation, ***Soviet Mathematics Doklady***, vol. 32 (1985), pp. 403–405.

[2] ALESSANDRO BERARDUCCI, *The interpretability logic of Peano arithmetic* (preliminary version), Manuscript, 1988.

[3] ———, Σ_n^0 *interpretations of modal logic*, ***Bollettino dell'Unione Matematica Italiana***, ser. 7, vol. 3-A (1989), pp. 177–184.

[4] GEORGE BOOLOS, ***The unprovability of consistency***, Cambridge University Press, Cambridge, 1979.

[5] DICK DE JONGH and FRANK VELTMAN, *Provability logics for relative interpretability*, ***Proceedings of Heyting '88*** (to appear).

[6] SOLOMON FEFERMAN, *Arithmetization of metamathematics in a general setting*, ***Fundamenta Mathematicae***, vol. 49 (1960), pp. 33–92.

[7] HARVEY FRIEDMAN, *Translatability and relative consistency*. II, manuscript.

[8] PER LINDSTRÖM, *Some results on interpretability*, ***Proceedings from 5th Scandinavian logic symposium*** (F. V. Jensen et al., editors), Aalborg University Press, Aalborg, 1979, pp. 329–361.

[9] ———, *Provability and interpretability in theories containing arithmetic*, ***Atti degli incontri di logica matematica (Siena, 1983, 1984***; C. Bernardi and P. Pagli, editors), Vol. 2, Università di Siena, Siena, 1985, pp. 431–451.

[10] RICHARD MONTAGUE, *Semantical closure and non-finite axiomatizability*. I, ***Infinitistic methods (proceedings of the symposium on foundations of mathematics, Warsaw, 1959)***, PWN, Warsaw, and Pergamon Press, Oxford, 1961, pp. 45–69.

[11] PAVEL PUDLÁK, *Cuts, consistency statements and interpretability*, this JOURNAL, vol. 50 (1985), pp. 423–441.

[12] HELMUT SCHWICHTENBERG, *Some applications of cut-elimination*, ***Handbook of mathematical logic***, North-Holland, Amsterdam, 1977, pp. 867–895.

[13] V. Yu. SHAVRUKOV, *Logic of relative interpretability over Peano arithmetic*, Preprint No. 5, Steklov Mathematical Institute, Academy of Sciences of the USSR, Moscow, December 1988.

[14] C. SMORYŃSKI, *Nonstandard models and related developments in the work of Harvey Friedman*, ***Harvey Friedman's research on the foundations of mathematics***, North-Holland, Amsterdam, 1985, pp. 212–229.

[15] ———, ***Self-reference and modal logic***, Springer-Verlag, Berlin, 1985.

[16] ROBERT M. SOLOVAY, *Provability interpretations of modal logic*, ***Israel Journal of Mathematics***, vol. 25 (1976), pp. 287–304.

[17] V. A. VARDANYAN, *Arithmetic complexity of predicate logics of provability and their fragments*, ***Doklady Akademii Nauk SSSR***, vol. 288 (1986), pp. 11–14; English translation in ***Soviet Mathematics Doklady***, vol. 33 (1986), pp. 569–572.

[18] ALBERT VISSER, *Preliminary notes on interpretability logic*, Logic group preprint series, no. 29, Department of Philosophy, University of Utrecht, Utrecht, 1988.

[19] ———, *Interpretability logic*, Logic group preprint series, no. 40, Department of Philosophy, University of Utrecht, Utrecht, 1988.

VIA STIBBERT 33-A
50134 FIRENZE, ITALY