Reasoning About an Ordering of Mathematical Theories: Modal System ILM is the Logic of Interpretability Over PA

Adam P. Lesnikowski

Submitted in partial fulfillment
of the requirements for the degree of
Bachelor of Arts, with Honors,
in Philosophy and Mathematics,
at Harvard University.

April 3, 2009

Contents

Chapter 1

Introduction

This thesis concerns interpretability, a concept in logic describing how mathematical theories relate to one another. In particular it is meant to clarify the notion that two theories are talking about the same thing, even though superficially they may appear to be very different. The central questions I address here are: Can the interpretability relation be made precise? And if so, is there a simple deductive framework to reason about it? The answer to both these questions, perhaps surprisingly, will turn out to be "yes."

Interpretations are ubiquitous in mathematics. Problems are often solved in a generalized or entirely new settings. Interpretations are implicitly used to justify that we are still talking about the same problem and hence that the proof transfers. This is especially true when the new setting is dramatically different from the original one. We can prove propositions of Euclidean geometry (for example that two triangles with three equal sides are similar) by proving their translations in the coordinate plane of analytic geometry. The grounds for this approach is that there exists an interpretation between the two theories. Points are translated into ordered pairs of real numbers, and geometric objects such as ellipses into sets of ordered pairs. Even though Euclidean points and pairs of real numbers are very different things, the interpretation allows us to treat them mathematically as if they were the same. One remarkable application is Alfred Tarski's proof of the consistency and decidability of Euclidean geometry by demonstrating the decidability of real closed fields together with a translation of geometric statements into the language of real closed fields.

Mathematical logic can formalize the admittedly imprecise notions of interpretability just described. One theory, such as set theory, *interprets* another, such as number theory, when there exists a map sending the symbols of the first theory to the symbols of the second theory respecting logical form such that the translations of all theorems of the first are theorems of the second. An interpretation can be said to embed a logical image of the base theory into the target theory. Interpretability then formalizes that one theory is at least as strong as another, in the same way that provability formalizes that one set of statements is a proof of a conclusion.

The concept of interpretability has figured prominently in the foundations of mathematics. At the end of the eighteenth century, Gottlob Frege sought to describe arithmetic in purely logical terms. He thought that he showed something philosophically meaningful in his reduction of arithmetic to logic: the nature of numbers is the nature of logic, and that the ultimate justification for arithmetic truths is precisely the justification that we have for logic. In modern terms we can say Frege was trying to show that logic interprets arithmetic. David Hilbert, a few decades later, hoped to show that large portions of mathematics were "ideal," useful for streamlining proofs, but ultimately translatable into purely finitary number-theoretic statements. His chief concern was the epistemic problem of knowledge of infinite objects. The question could be defused though by showing that infinite objects were not necessary in mathematics. Moreover the instrumentally useful of ideal objects in mathematics could be justified by a translation of ideal elements into finitary ones. At the heart of both Frege's and Hilbert's foundational projects was the idea that if one can show that a theory is interpretable in a second theory, then the first theory is (mathematically, epistemologically, metaphysically) reduced to the second. In this way they sought a philosophical grip on suspect theories like analysis by interpreting them in a philosophically sound theory like finitary arithmetic.

Recently the study of interpretability for its own sake has yielded several important contributions to mathematical logic. Interpretability is a proper generalization of the formal provability predicate. To say that we can prove a statement is equivalent to saying that its negation interprets a contradiction. It is no surprise then that interpretability has helped illuminate Kurt Gödel's celebrated 1931 Incompleteness Theorems. Interpretability turns out to be one of the best frameworks to study incompleteness phenomena. The relation allows us to arrange naturally arising theories in a hierarchy, where one theory is above a second theory if and only if the second theory can be interpreted into the first theory. In this hierarchy, all complete theories are arranged below all incomplete theories. By studying theories near the threshold point of incompleteness, we can probe what features theories have that make them expressive enough so that they cannot prove everything expressible in their language, that is, incomplete.

Interpretability allows us to study much stronger theories as well. We have that if a theory can be interpreted into another, then the consistency of the interpreting theory implies the consistency of the interpreted theory. Working with the assumption that relative consistency is a suitable measure of strength, interpretability is also then a reliable indicator of theory strength. The interpretability hierarchy today is used to gauge the strength of newly proposed axiomatic systems that include large cardinal axioms. These additional assumptions provide theories the strength to decide statements not settled by standard axiomatizations of set theory. This ordering disregards differences that arise from presentation rather than content, and so is an ordering "up to isomorphism." Indeed this approach allows logic, as other areas of mathematics successfully do, to deal with general structure rather than particular appearances, providing the means to prove general statements about its objects of study. The philosophical payoff of the interpretability hierarchy then is that we can measure, in a rigorous way, the strength of these newly proposed axioms.

Nonetheless important foundational questions about interpretability remain unanswered. For instance, it has only been an empirical observation that every naturally occurring formal theory so far studied is linearly ordered by the hierarchy of interpretability. We currently have no principled way to say why this is so. One productive line of research has been towards a modal logic of interpretability. This is the search for a relatively simple logic that captures propositional reasoning about interpretability in. Questions include: Is interpretability between theories transitive? Can there be so-called Orey statements ϕ such that a theory interprets itself plus ϕ while *also* interpreting itself plus not ϕ? When dealing with interpretability over Peano Arithmetic, the answer to both will be positive. A particular modal system ILM answers these and many other related questions. In fact ILM will be seen to be *the* modal logic of interpretability, exactly capturing the correct propositional statements involving interpretability.

At first glance this reduction seems impossible. To proof-theoretically formalize the interpretability relation involves an astronomically large number of symbols, while the modal system can be written in just nine axioms and two rules of inference, a mere few hundred symbols. Yet it can be done. In this thesis, I will describe how the transmutation is carried out.

Chapter 1 is this introduction. Chapter 2 introduces interpretability and the necessary metamathematical preliminaries. Chapter 3 defines the modal system ILM and provides a sound and complete semantics for the logic. Chapter 4 brings together interpretability with the modal system. In this final chapter, I present the main proof of this work, the arithmetical soundness and completeness of ILM with respect to interpretability over PA.

Chapter 2

Preliminaries

In this chapter, we will introduce a number of preliminaries that will be needed in the presentation to follow. In particular, we will introduce both the historical and modern notions of interpretability, along with basic results about formalized metamathematics and formalized model theory.

2.1 Original Notion of Interpretability

In this section, we will define the notions of interpretability and weak interpretability as they first appeared in the literature in [10]. We closely follow the presentation in that work. We begin with the concept of a possible definition, introduced the following example:

Example 1. Suppose that we have a theory T in a language $\mathcal{L}_T$ which does not contain a certain non-logical constant, say the binary predicate $<$ (the less-than equal sign). A *possible definition* of $<$ in T is any sentence of the following form:

$$\forall x \forall y (x < y \leftrightarrow \Phi) \tag{2.1}$$

where Φ is a formula of T. Of course (2.1) is not a formula T, as we assumed T does not contain the constant $<$. But it is a formula, in fact a sentence, of every extension of T that contains $<$ as a constant. In this way we have defined $<$ for extensions of T in terms of a formula expressible in T.

Definition. Let T_1 and T_2 be any two theories in languages $\mathcal{L}_1$ and $\mathcal{L}_2$. First assume that the two languages share no non-logical constants. Then

T_1 *interprets* T_2 if we can extend T_1 by including in its set of valid sentences some possible definitions of all of the non-logical constants of T_2 so that the resulting extension is also an extension of T_2. Equivalently we say that T_2 is *interpretable* in T_1. More precisely, T_1 interprets T_2 if and only if there is a theory T and a set of sentences D such that:

1. T is a common extension of T_1 and T_2 and every constant of T is either a constant of T_1 or of T_2;

2. D is a recursive set of sentences which are all valid in T and which are possible definitions in T_1 of the non-logical constants of T_2;

3. Every non-logical constant of T_2 occurs in exactly one sentence of D;

4. Every valid sentence (in T) is derivable in T from T_1 along with the set D.

In general, when T_1 and T_2 share some non-logical constants, we first replace the non-logical constants of T_2 with new constants not occurring in T_1 (different symbols by different symbols) to obtain T_2'. This is done so that T_2 and T_2' have no differences in their structures beyond the ways they are presented. Then if T_2' is interpretable in T_1, we say that T_2 is interpretable in T_1 as well.

Definition. A theory T_1 *weakly interprets* T_2 if there is some consistent extension of T_1 which has the same constants as T_1 and interprets T_2.

Definition. Let T be a theory and P be a one-place predicate of T. For every formula Φ of T replace every subformula of the form $\forall x\Psi$ or $\exists x\Psi$ by the expressions:

$$\forall x(Px \to \Psi) \tag{2.2}$$

$$\exists x(Px \wedge \Psi) \tag{2.3}$$

The resulting formula Φ^P is said to be the *relativization of* Φ *to* P. If we fix P and for each Φ in T take its relativization Φ^P, then the collection we obtain is the new theory T^P, the *relativization of* T *to* P. The set of all constants of T^P will be all the constants of T and of the predicate P. We stipulate that a sentence is valid in T^P if and only if it is derivable from the set of sentences Φ^P obtained by relativizing the set of valid sentences Φ of T by P.

Definition. Let T_1 and T_2 be any two theories. Then T_1 *relatively interprets*, or *weakly relatively interprets* T_2 if and only if there is a predicate P that does not occur in T_2 and T_1 interprets, or weakly interprets T_2^P in the sense of Definitions (2.1) and (2.1).

A relative interpretation can be thought of as an interpretation in which an interpretation is provided for not only of the non-logical constants but also for quantifiers appearing in formulas.

Definition. A theory T is called *decidable* if the set of all its valid sentences is recursive, otherwise it is called *undecidable.* A undecidable theory T is called *essentially undecidable* if every consistent extension of T that has the same constants as T is also undecidable.

These notions of interpretability were historically introduced to prove the decidability or undecidability of some theory. In particular, we can show that a theory is undecidable if we can interpret an already known undecidable theory in it. But for some theories providing an interpretation or weak interpretation in the usual sense is either difficult or impossible. There are many cases though in which a theory T_2 can easily be show to be essentially undecidable and relatively interpretable or weakly relatively interpretable in a given theory T_1. We have results that relative and weak relative interpretations also provide us our desired decidability results. Hence these notions greatly expand our ability to reason about the undecidability of theories. We will now turn to a closer study of what we introduced as relative interpretability. First though, we will need to introduce some facts about formalizing mathematics.

2.2 Peano Arithmetic: the Setting

Peano Arithmetic (PA) is a first-order theory intended to formalize number theory. It has the non-logical symbols $0, S, +, \times$ for zero, successor, addition, and multiplication respectively. The terms (or collection of symbols) '0', '$S(0)$', '$S(S(0))$', ... are called *numerals.* The n^{th} numeral will be denoted by $\mathbf{n}$.

By ω we denote the set of natural numbers, also identifiable with the first infinite ordinal. The structure $\mathbb{N}$ whose underlying set is ω and equipped with the operations successor, addition, and multiplication will be called the standard model (or intended interpretation) of PA. The interpretation of the numeral $\mathbf{n}$ in this model is the natural number n.

We call any finite sequence of symbols from the alphabet of PA a *syntactic object*. Formulas, terms, and proofs are all syntactic objects. With every syntactic object t we bijectively associate in an effective way a numeral $\ulcorner t \urcorner$ called the *Gödel number of* t.

Definition. The T-formula $\alpha(x_1, \ldots, x_k)$ *defines* the number-theoretic relation R if for all $n_1, \ldots n_k \in \omega$,

$$\mathbb{N} \models \alpha(\mathbf{n_1}, \ldots, \mathbf{n_k}) \iff R(n_1, \ldots, n_k) \tag{2.4}$$

The T-formula $\alpha(x_1, \ldots, x_k)$ *numerates* the number-theoretic relation R if for all $n_1, \ldots n_k \in \omega$,

$$T \vdash \alpha(\mathbf{n_1}, \ldots, \mathbf{n_k}) \iff R(n_1, \ldots, n_k) \tag{2.5}$$

The T-formula $\alpha(x_1, \ldots, x_k)$ *binumerates* the number-theoretic relation R if for all $n_1, \ldots n_k \in \omega$, if it numerates R and in addition,

$$T \vdash \neg\alpha(\mathbf{n_1}, \ldots, \mathbf{n_k}) \iff \neg R(n_1, \ldots, n_k) \tag{2.6}$$

In the following we state without proof many results which we assume the reader to have had some exposure to. Many of these proofs, including the next theorem, can be found in [5].

Binumerability Theorem 2. *For every primitive recursive number theoretic relation $R(x_1, \ldots, x_k)$, there is a* PA*-formula $\alpha(x_1, \ldots, x_k)$ which binumerates R.*

In the proof of Theorem 2, the formula α is explicitly constructed from the number-theoretic relation R. Any formula so obtained from a primitive recursive relation will be called a *primitive recursive* formula.

Definition. We can group formulas into classes designated as either Σ_n or Π_n. These are defined recursively as follows:

(i) $\Sigma_0 = \Pi_0 =$ the class of formulas with bounded quantification (also notated Δ_0),

(ii) Σ^0_{n+1} is the class of all formulas obtained by prefixing some number of existential quantifiers in front of a Π_1-formula, and

(iii) Π^0_{n+1} is the class of all formulas obtained by prefixing some number of universal quantifiers in front of a Σ_1-formula.

2.3 Modern Interpretability

This section details the modern notion of interpretability that will be studied through the rest of this paper. The first full treatment appeared in [4], where the idea of relative interpretability (defined above) was studied in greater detail. The presentation of this section follows [1] and [8].

A *theory* (unless otherwise noted) will denote an axiomatic theory formulated in first-order logic with equality in a finite language and with a recursively enumerable set of axioms. What we have in mind for an interpretation is a map i of the language of S to the language of T so that i commutes with boolean connectives, and if $S \vdash \phi$, then $T \vdash i(\phi)$. If such an interpretation exists, we will sometimes write $S \leq T$. For example, if $S \vdash \phi \vee \psi$ and $S \leq T$, then $T \vdash i(\phi) \vee i(\psi)$. From this it follows if $T \vdash S$, then $S \leq T$. This map can be thought of as an embedding of a logical image of the the theory S into the theory T. Given this syntactic definition in mind, we now state a model-theoretic, semantic version that we will find easier to work with.

Definition. An *interpretation* f of a $\mathcal{L}_S$-theory S in into a $\mathcal{L}_T$-theory T is a set of $\mathcal{L}_T$-formulas which define, for every model $\mathcal{M}$ of T, the universe, set of relations (except equality), and graphs of the functions of a $\mathcal{L}_S$-model $\mathcal{M}^f$ of S. Then $\mathcal{M}^f$ is called the *interpreted structure* or the *interpreted model.* We stipulate that the equality symbol will always be interpreted as the identity relation. The underlying set of $\mathcal{M}^f$ will be $\{\, a \in \mathcal{M} \mid \mathcal{M} \models \delta(a) \,\}$, where $\delta(x)$ is the formula of the interpretation defining the universe.

An interpretation will induce a map i which assigns to every $\mathcal{L}_S$ formula $\phi(x_1, \ldots, x_n)$ a $\mathcal{L}_T$ formula $\phi(x_1, \ldots, x_n)^f$ (with the same free variables) so that for all $a_1, \ldots, a_n \in \mathcal{M}^f$,

$$\mathcal{M}^f \models \phi(a_1, \ldots, a_n) \Leftrightarrow \mathcal{M} \models \phi(a_1, \ldots, a_n)^f \tag{2.7}$$

2.4 Formalized Metamathematics

Diagonal Lemma 3. *For every formula $\phi(\vec{x}, y)$ there is a formula $\alpha(\vec{x})$ such that* $\mathrm{PA} \vdash \forall \vec{x}(\alpha(\vec{x}) \leftrightarrow \phi(\vec{x}, \ulcorner \alpha \urcorner))$.

The formula guaranteed by Lemma 3 is defined relative to ϕ in a primitive recursive way. In particular, if ϕ is primitive recursive, then α will at worst be provably equivalent to a primitive recursive formula.

Given an arithmetically axiomatized theory T in the language of PA, we can associate to T a *proof predicate*. This is a formula $\mathrm{Prf}_T(x, y)$ that is a formalization of the statement that 'x is a proof of y in T'.

Note that $\mathrm{Prf}_T(x, y)$ depends not only on the set of axioms of T, but also on the way they are presented, namely the formula $\tau(x)$ that defines that set of axioms. A less ambiguous notation would then be $\mathrm{Prf}_{T,\tau}(x, y)$. For brevity by τ we will denote both the theory and a formula presenting its axioms.

The formula $\mathrm{Pr}_\tau(x)$ is defined as $\exists x(\mathrm{Prf}(x, y))$ and so is a formalization of 'y is a theorem of τ'. The formula $\mathrm{Con}(\tau)$ is defined as $\neg \mathrm{Pr}_\tau(\ulcorner 0 = 1 \urcorner)$ and so is a formalization of 'τ is consistent'.

Notation. If T is a theory and ϕ is a sentence, let $T + \phi$ denote adding ϕ as an axiom to T.

Notation. Let $T \upharpoonright n$ be the subtheory of T axiomatized by the axioms of T with Gödel numbers $< n$. In particular let T_k stand for $T \upharpoonright k$. If $\sigma(x)$ is a formula, let $\sigma(x) \upharpoonright y$ stand for the formula $\sigma(x) \wedge x < y$. In particular if $\sigma(x)$ defines a theory T, then $\sigma(x) \upharpoonright \mathbf{n}$ will define the theory T_n. If T is a finite theory, let $[T](x)$ be the canonical formula defining T obtained by taking the disjunction of all formulas of the form $x = \mathbf{n}$ where n is the Gödel number of an axiom of T.

Provable Σ_1-Completeness of PA 4. PA $\vdash$ *'for every ϕ, if ϕ is a true Σ_1-sentence, then* PA $\vdash \phi$*'.*

Provable Σ_n-Soundness of PA$_k$ 5. PA $\vdash$ *"for every k, n,* PA *proves 'for every Σ_n-sentence ϕ, if* PA$_k \vdash \phi$*, then ϕ is true' ".*

In other words Theorem 5 states that PA proves, for any n, the Σ_n-soundness of every finite fragment of itself. This result will sometimes be called 'Reflection'. From this also follows:

Corollary 6. *For all k,* PA *proves the consistency of* PA$_k$.

Gödel's Completeness Theorem states that every consistent theory has a model. This can be stated in PA in the following way:

Formalized Gödel's Completeness Theorem 7. *Let T be a theory that contains* PA*, and S a first-order theory. Suppose that $T \vdash \mathrm{Con}(\tau(x))$, where $\tau(x)$ numerates S in T.*[1] *Then T interprets S.*

[1] i.e. $T \vdash \tau(\mathbf{x}) \iff R(x)$, where $R(x)$ here holds if and only if x is the Gödel number of a theorem of S.

Here, the formalized version of the original hypothesis: 'S is consistent' becomes 'the PA-extension T proves $\mathrm{Con}(\tau)$', where $\mathrm{Con}(\tau)$ is the formalized statement of S's consistency relative to a presentation $\tau(x)$ of S. The formalized version of the original conclusion: 'S has a model' becomes 'T interprets S'. One way then to think about an interpretation is as a syntactical procedure to build in the interpreting theory a formalized model of the interpreted theory.

It is a classic result that every recursively enumerable set can be defined by a Σ_1 formula, and conversely that every Σ_1 formula defines a recursively enumerable set. A theorem by William Craig states that every theory that has a recursively enumerable set of axioms also has a primitive recursive set of axioms. Soloman Feferman proved the following formalized version of Craig's result:

Formalized Craig's Theorem 8. *Let $\xi(x)$ be a Σ_1-formula. Then there is a primitive recursive formula $\alpha(x)$ such that* $\mathrm{PA} \vdash \mathrm{Pr}_\xi(x) \leftrightarrow \mathrm{Pr}_\alpha(x)$.

The following theorem of Stephen Orey will be important for this presentation. It equates interpretability with the interpreting theory proving the consistency of all finite sub-theories of the interpreted theory:

Orey's Theorem 9. *Let T be a theory that contains* PA*, and let S be a theory with a recursively enumerable set of axioms. If $T \vdash \mathrm{Con}(S')$ for every finite sub-theory S' of S, then T interprets S.*

Proof. ($\Rightarrow$) By Theorem 8, we can assume that S is axiomatized by a primitive recursive formula $\sigma_0(x)$. Extend this to $\sigma(x)$, a primitive recursive formula that binumerates the theory S in PA. Then the hypothesis of the theorem can be restated as: $\forall k(T \vdash \mathrm{Con}(\sigma \restriction k))$. Now let:

$$\sigma^*(x) = \sigma(x) \wedge \mathrm{Con}(\sigma \restriction x+1) \tag{2.8}$$

This is done so $\sigma^*(x)$ defines (in ω) the consistent subtheories of S of the form $\sigma(x) \restriction y$. By a formalization in PA of Gödel's Compactness Theorem, we get:

$$\mathrm{PA} \vdash \mathrm{Con}(\sigma^*(x)) \tag{2.9}$$

From the Theorem's hypothesis and equation (2.8), if T is consistent, then $\sigma^*(x)$ binumerates (and hence numerates) S in T. Then equation (2.9), and

Theorem 7 give the desired result that T interprets S. On the other hand in T is inconsistent, then T interprets any theory, including S, and we are done.

($\Leftarrow$) Suppose T interprets S. We must show that for any finite subtheory S' of S, $T \vdash \mathrm{Con}(S')$. First we obtain a finite subtheory T' of T so that T' interprets S' by taking the inverse of the interpretation map on all of S''s axioms (remember that an interpretation is a bijection between S and some subtheory of T, possibly T itself). It is a fact that for finitely axiomatized theories, the interpretability relation can be formalized in PA as a Σ_1-assertion. By Theorem 4:

$$\mathrm{PA} \vdash \text{`}T' \text{ interprets } S'\text{'} \tag{2.10}$$

As T' interprets S' the interpretation of anything (including a contradiction) provable in S' is provable in T'. Because the interpretation of $\perp$ is always $\perp$, if S' is inconsistent, then T' is inconsistent. Stating the contrapositive of this in PA we get:

$$\mathrm{PA} \vdash \mathrm{Con}([T']) \rightarrow \mathrm{Con}([S']) \tag{2.11}$$

As T extends PA, by Corollary 6, $T \vdash \mathrm{Con}([T'])$. Combining this with equation (2.11) gives the desired conclusion $T \vdash \mathrm{Con}([S'])$, where $[S']$ is the canonical formula defining the finitely axiomatizable theory S'.

□

2.5 Model Theory in PA

The second-order theory ACA_0 can formalize some model theoretic notions. Moreover, it can be shown that ACA_0 is a conservative extension of PA. This means that if we have a formula in the language of PA, whatever ACA_0 proves can be proved in PA as well. Hence, we can use ACA_0 as a tool to prove that things hold in PA. The advantage is that we can formalize more complicated notions in ACA_0, hopefully yielding quicker proofs in ACA_0 of statements of PA.

Definition. The language of ACA_0 properly extends PA. In addition to the usual numerical variables $x, y, z, \ldots$ of PA, set variables $X, Y, Z, \ldots$ are included in the language of ACA_0. The language of ACA_0 also includes a binary relation '$\in$ ' whose intended meaning when '$x \in X$' is that the number x is a member of the set X. The axioms of ACA_0 are:

1. 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. (Arithmetical Comprehension Schema) Let $\phi(x)$ be any formula containing no bound set variables where x is not free, and let X be a set variable not occurring in ϕ. Then there is an axiom which asserts the universal closure of: $\exists X \forall x(x \in X \leftrightarrow \phi(x))$.

In words, axiom (2) allows us to conclude the general statement $\forall x(x \in X)$ whose intended meaning is 'all numbers are in the set X' from the specified induction clauses. Axiom (3) guarantees that for each formula satisfying the specified conditions there is a set that contains exactly those elements for which the formula holds. Note in axiom (3) there may be bound numerical variables and free numerical or set variables.

Definition. Suppose that we have two $\mathcal{L}_{\mathrm{PA}}$-structures $\mathcal{M}$ and $\mathcal{N}$ such that $\mathcal{M}, \mathcal{N} \models$ PA, $\mathcal{N}$ a substructure of $\mathcal{M}$. Then $\mathcal{N}$ is an *initial segment* of $\mathcal{M}$, or $\mathcal{M}$ is an *end-extension* of $\mathcal{N}$, iff

$$\text{for all } x \in \mathcal{N}, \text{for all } y \in \mathcal{M}(\mathcal{M} \models y < x \implies y \in \mathcal{N}) \tag{2.12}$$

This situation will be denoted in symbols by $\mathcal{N} \subseteq_e \mathcal{M}$. N is a *proper* initial segment if in addition $M \neq N$.

Theorem 10. *Let $\mathcal{Y}, \mathcal{Z}$ be models of* PA *with $\mathcal{Z}$ an end-extension of $\mathcal{Y}$. Then for all Σ_1 sentences A, $\mathcal{Y} \models A \implies \mathcal{Z} \models A$.*

Theorem 11. *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 equality).*

Proof. Let $G(x, y)$ be the formula of PA that formalizes the following: there is a finite sequence $u = \langle u_0, u_1, \cdots, u_x \rangle$ such that $(u_0 = 0)^f$, $u_x = y$, and for all $i \leq x$, $(S(u_1) = u_{i+1})^f$. 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 $\mathcal{M}^f$ (the verification is left to the reader that g preserves $+$ and $\times$and so is an embedding). The sentence $\forall u, x(u < x+1 \to u < x \vee u = x)$ is a theorem of PA, and hence holds in the models $\mathcal{M}$ and $\mathcal{M}^f$. Specifically, $\mathcal{M} \models (u < g(x+1) \to u < g(x) \vee u = g(x))^f$, by substitution of $g(x)$ for x above. It follows then (by induction on $x \in \mathcal{M}$) that, for every $u \in \mathcal{M}$ such that $\mathcal{M} \models (u < g(x))^f$, there is a $y < x \in \mathcal{M}$ such that $\mathcal{M} \models (u = g(y))^f$. This means g embeds $\mathcal{M}$ as an initial segment of $\mathcal{M}^f$ □

Theorem 12. *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. ($\Rightarrow$) This follows from the previous theorem. ($\Leftarrow$) Proof by contrapositive. If PA $+ \alpha$ does not interpret PA $+ \beta$, then by Orey's theorem, there exits some k such that PA $+ \alpha \nvdash \mathrm{Con}(\mathrm{PA}_k + \beta)$. Consider a model of PA$+\alpha$ in which $\neg\mathrm{Con}(\mathrm{PA}_k+\beta)$ holds. Being a Σ_1-assertion, $\neg\mathrm{Con}(\mathrm{PA}_k+\beta)$ then must hold in every end-extension $\mathcal{N}$ of $\mathcal{M}$ (by a fact of such assertions in end-models). Then $\mathcal{N}$ can't be a model of PA $+ \beta$. If it were, we would have that both $\mathcal{N} \models \neg\mathrm{Con}(\mathrm{PA}_k + \beta)$, and by reflection that for all k, $\mathcal{N} \models \mathrm{Con}(\mathrm{PA}_k + \beta)$. This is absurd. So no end-extension $\mathcal{N}$ of the model $\mathcal{M}$ of PA $+ \alpha$ models PA $+ \beta$. □

If $\mathcal{M}$ is a model of PA and $\forall k \mathcal{M} \models \mathrm{Con}(\mathrm{PA}_k+\phi)$, then Orey's Theorem 9 guarantees that the theory of $\mathcal{M}$ interprets the theory PA $+ \phi$. Furthermore the interpreted structure $\mathcal{M}^f$ is an end-extension of $\mathcal{M}$ which models PA$+\phi$. This construction can be formalized in ACA$_0$ (even if ϕ contains non-standard elements as calculated by $\mathcal{M}$) to give the following:

Theorem 13. ACA$_0 \vdash$ *'if $\mathcal{M}$ is a model of* PA *with $m \in \mathcal{M}$, $\phi(m)$ a formula, and $\forall k[\mathcal{M} \models \mathrm{Con}(\mathrm{PA}_k + \phi(\boldsymbol{m}))]$, then there is an end-extension $\mathcal{N}$ of $\mathcal{M}$ such that $\mathcal{N} \models$* PA $+ \phi(\boldsymbol{m})$*'.*

Chapter 3

Semantics for ILM

Our goal in this chapter will be to formalize interpretability. This will be done by providing a modal deductive system with a binary operator whose intended interpretation is that one sentence interprets another over the base theory PA. We will also prove in this chapter a model completeness theorem for this system. This completeness theory will state that if the system disproves a statement, then then there will be a structure of a particular kind that satisfies the negation of the statement. Together with the fact that if the system proves a statement, then all structures of this particular kind satisfy the statement, we will have a complete semantics for the deductive system.

3.1 Axiomatization of ILM

The language $\mathcal{L}(\rhd)$ of interpretability logic consists of a set of propositional variables $p_1, p_2, \ldots$, connectives $\vee, \wedge, \rightarrow, \leftrightarrow, \neg, \bot$, a unary operator $\Box$, and a binary operator $\rhd$. Note we can consider some of the connectives as abbreviations of combinations of other connectives, e.g. $A \wedge B$ as $\neg(\neg A \vee \neg B)$ and $\bot$ as any contradiction. Let $\Diamond$ be an abbreviation of $\neg\Box\neg$.

Definition. The modal theory ILM in $\mathcal{L}(\rhd)$ is axiomatized by all the tautologies plus the following axiom schemes:

1. $\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
2. $\Box(\Box A \rightarrow A) \rightarrow \Box A$
3. $\Box A \rightarrow \Box\Box A$

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:

(i) if $\vdash A$ and $\vdash A \to B$, then $\vdash B$ (modus ponens), and

(ii) if $\vdash A$ then $\vdash \Box A$ (necessitation).

Note that restricting this axiomatization to the first three axioms and the two rules of inferences results in the well-known modal system GL, also known as the *logic of provability*. In this system, the intended interpretation of $\Box A$ is ‘PA proves A’.

It is also a well-known fact that axiom 3 is in fact derivable from axioms 1,2 along with the rules of inferences, though the derivation is nontrivial. We will nontheless include it as an axiom for the sake of presentation.

For more information on GL, including the derivation of axiom 3, see [2].

3.2 Frames

Definition. A *frame* $\mathbf{W}$ is an ordered pair $\langle W, R\rangle$ consisting of a nonempty set W and a binary relation R on W.

Definition. We call $\langle W, R\rangle$ *finite* if an only if the set W is finite.

Definition. A relation R on a set Y is called *converse well-founded*, or *c.w.f.* for short, if for every nonempty set $X \subset Y$, there is an R-greatest element of X, that is, an element $m \in X$ such that xRm for no $x \in X$.

Proposition 14. *If a relation R is converse well-founded, then it is irreflexive.*

Proof. If Y is empty, then any relation R on Y will be vacuously irreflexive. Without loss of generality, Y is non-empty. Assume for contradiction that R on Y is converse well-founded and reflexive. Let x be an element of Y, xRx, and let $X = \{x\}$. But then there is no R greatest element of X. Contradiction. R is irreflexive. □

Definition. We call a frame $\mathbf{W} = \langle W, R\rangle$ *transitive* if and only if the relation R is transitive, and call the frame *converse well-founded* if and only if the relation R is converse well-founded.

Remark 15. Elements of W will occasionally be called 'possible worlds,' 'worlds,' or 'nodes'. The fact that a world w stands in the relation R to a world x, so that wRx, will sometimes be denoted by saying that world w 'sees' world x. The terminology 'words' is due to the original development of these semantics as a formalization of the notions of metaphysical necessity and possibility.

Definition. A frame **W** will be called a GL-*frame* if and only if it is transitive and converse well-founded.

3.3 ILM-Models

Definition. An ILM-*frame* is a GL-frame $\langle W, R\rangle$ with, for each $w \in W$, an addition relation S_w, which has the following properties:

(i) S_w is a relation on $w^{\uparrow} = \{w' \in W | wRw'\}$,

(ii) S_w is reflexive and transitive,

(iii) if $w', w'' \in w^{\uparrow}$ and $w'Rw''$, then $w'S_w w''$.

(iv) if there is a w such that $xS_w yRz$, then xRz.

Notation. We will sometimes write S for $\{S_w | w \in W\}$ when we are dealing with ILM-models.

Definition. A *forcing relation* $\Vdash$ is defined as a relation between worlds and propositional variables. The relation for a particular model is sometimes called a *valuation* map as it assigns a truth value to every propositional variable at every world in the model. The relation is extended to one between worlds and formulas by the following stipulations, so that for all worlds x, y, z and formulas A, B:

1. $x \Vdash \neg A \iff x \nVdash A$,

2. $x \Vdash A \vee B \iff x \Vdash A$ or $x \Vdash B$,
   and similarly for the other connectives,

3. $x \Vdash \Box A \iff \forall x(xRy \Rightarrow y \Vdash A)$,

4. $x \Vdash A \rhd B \iff \forall y(xRy \wedge y \models A \Rightarrow \exists z(yS_x z \wedge z \models B))$.

Definition. An ILM-*model* is given by an ILM-frame together with a forcing relation $\Vdash$.

Notation. If F is a frame, then we write $F \models A$ iff $F = \langle W, R, S\rangle$ and $w \Vdash A$ for every $w \in W$ and every $\Vdash$ on F. If $\mathcal{K}$ is a class of frames, we write $\mathcal{K} \models A$ iff $F \models A$ for each $F \in \mathcal{K}$. $\mathcal{K}_{FM}$ will denote the class of finite ILM-frames.

3.4 Model Completeness

We now state and prove the main theorem of this chapter. The original result was proved by Dick de Jongh and Frank Veltman in [3].

Finite Model Soundness and Completeness Theorem for ILM 16. *For each A, $\vdash_{\mathrm{ILM}} A$ if and only if $\mathcal{K}_{FM} \models A$.*

Proof. The ($\Rightarrow$) direction is called the *soundness of* ILM *with respect to finite* ILM-*models*, or *finite model soundness of* ILM for short. This direction routine to verify, and amounts to showing that all the axioms of ILM are valid at every node of every finite ILM-model, and that each rule of inference preserves validity in every finite ILM-model (in fact these holds for all ILM-frames, finite or not).

The ($\Leftarrow$) direction is called the *completeness of* ILM *with respect to finite* ILM-*models*, or *finite model completeness of* ILM for short. In particular, we will show that if $\vdash_{\mathrm{ILM}} \neg A$, then there exists a (finite) ILM-model $\mathbf{W}$, with root b, such that $b \Vdash \neg A$.

To prove this direction, we will first develop the notions of adequate sets of formulas and of critical successors of these sets.

3.4.1 Adequate Sets

Definition. An *adequate set* of formulas in the language $L(\rhd)$ 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 in an antecedent or a consequent of some $\rhd$-formula is in Φ, then $B \rhd C \in \Phi$.

6. If $B \rhd C, \Box D \in \Phi$, then there are formulas B', C', which are L-equivalent to $B \wedge \Box D$, $C \wedge \Box D$ receptively, such that $B' \rhd C' \in \Phi$.

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

Lemma 17. *Let U be some 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. Note X is infinite (consider $g_0 = f(A,A), g_{n+1} = f(A, f(n))$). Nonetheless X is included in only finitely many equivalence classes with respect to L-provable equivalence.*

Proof. The lemma follows from the special case in which $\bot \in U$ and all other elements of U are propositional variables. To prove this special case, proceed by induction on the cardinality n of U. Note that a formula C belongs to X iff $C = A \wedge \Box\neg D_1 \cdots \wedge \Box\neg D_n$, for $A \in U$, and all D_i's $\in X$ (possibly $n = 0$). Therefore to prove the lemma, it is enough to show that up to L-equivalence, there are only finitely many formulas of the form $\Box\neg D \in X$, say k many. Once this is established, we will have a bound on n as well, and so a bound on X up to L-provability (i.e. $\leq |U| \cdot 2^k$).

Base case: If $|U| = 1$, then $U = \bot$, and then every formula is L-equivalent to $\bot$, and we are done. Induction Case: So assume $|U| > 1$. We want to show that if lemma holds for $|U| = n$, then the lemma holds for $|U| = n+1$. Consider $D \in X$. Then, for some $n > 0$, D has the form $u \wedge E$, or $u \wedge \Box\neg F_1 \wedge \cdots \wedge \Box\neg F_n$, with $u \in U$, and all the F's in X. If $u = \bot$, then

$\Box\neg D$ is L-equivalent to $\Box\neg\perp$, (because $D =\perp \wedge E$). If $u \neq\perp$, then let E' be obtained by replacing all occurrences of the sentence variable u in E with $\perp$, so that $E' = \Box\neg F'_1 \wedge \cdots \wedge \Box\neg F'_n$. By the induction hypothesis, there are only finitely many choices for E' up to L-equivalence. (Warning: we cannot assume that $E' \in U$. Instead, apply the induction hypothesis to the 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 of L with respect to finite Kripke models (i.e. that $L \nvdash A \implies$ exists a finite L model V with root b s.t. b valuates A as false). First note that E, being a conjunction of $\Box$-formulas, is preserved upwards in Kripke models; that is, if $x \Vdash E$, and xRy, then $y \Vdash E$ (as R is transitive in Kripke models appropriate to L).

Now suppose that $L \nvdash \Box\neg(u \wedge E)$, so that exists a (finite) Kripke model with root b where $\Box\neg(u \wedge E)$ fails . 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 node (by the c.w.f. principle). So we must have that $x \Vdash u \wedge (\Box\neg u \wedge E)$. But this implies that $x \Vdash u \wedge E'$ (since every occurrence of u in E lies within the scope of a $\Box$-operator). Hence, $\Box\neg(u \wedge E')$ fails in this same Kripke model. The converse in completely similar. So $\Box\neg(u \wedge E)$ and $\Box\neg(u \wedge E')$ fail together in all (finite) Kripke models. □

Lemma 18. *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 $\perp \rhd \perp \in \Phi_1$. Define $n(A)$, the *pseudo-negation* of A, as: $n(A) = B$ if $A = \neg B$ and $n(A) = A$ otherwise. Let U_0 be the closure of Φ_1 under subformulas, and U the closure of U_0 under pseudo-negation. Then U is closed under both subformulas and pseudo-negations (i.e. we will not have a subformula appear in U that was not already in U_0). Let X be the union of an infinite sequence of sets of formulas $X_0, X_1, \ldots$, where $X_0 = U$. Let X_{n+1} be the union of X_n and the set of all formulas $F \wedge \Box\neg G$, $F, G \in X_n$, which are not L-equivalent to any formula in X_n. Clearly then X is closed (up to provable L-equivalence) under the operation $f(F, G) = F \wedge \Box\neg G$. By the previous lemma, X is still finite (i.e. f starts repeating itself). Now let:

$$\Phi_2 = U \cup \{B \rhd C | B, C \in X\} \cup \{\Box\neg A | A \in X\}. \tag{3.1}$$

Finally, let Φ_3 be the closure of Φ_2 under subformulas and pseudo-negations.

The claim now is: Φ_3 is a finite adequate set of formulas containing Φ_1. To prove this, we use the following facts which can be easily proved using induction on n (the indices of X):

1. If a formula $B \rhd C$ 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 $B \rhd C$, $\Box D$ 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 of U,

2. $B, C \in X$, or

3. $B \rhd C$ is a subformula of X, and therefore belongs to U, by the fact above.

Since $U \subseteq X$, and U is closed under subformulas, in all cases, $B, C \in X$. Similarly, from $\Box D \in \Phi_3$, it follows that $\Box D$ is a subformula of a formula in Φ_2. By similar reasoning as above, we have the either $D = \neg A$, for some $A \in X$, or $\Box D \in U$. Since U is closed under both subformulas and pseudo-negation, then in the second case, $\neg D \in U$, and so $\neg D \in X$. In either case, we have that $\neg D$ is L-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 B', C' in X. To get this, take $B' = f(B, x)$, where x is the formula in X that $\neg D$ is equivalent to (note that in the first case where $D = \neg A$, for some $A \in X$, both $B \wedge \Box D$ and $C \wedge \Box D$ need not be in X). This proves clause 6.

To prove clause 5, observe that if B, C is the antecedent or consequent of some $\rhd$-formula in Φ_3, then, reasoning as above, $B, C \in X$, and therefore $B \rhd C \in \Phi_2$, hence $B \rhd C \in \Phi_3$.

To prove clause 4, recall that $\Diamond = \neg\Box\neg$. Again, if $B \rhd C \in \Phi_3$, then $B, C \in X$, and hence $\Box\neg B, \Box\neg C \in \Phi_2$. But then $\neg\Box\neg B, \neg\Box\neg C \in \Phi_3$. As Φ_3 is closed under pseudo-negations, $\Diamond B, \Diamond C \in \Phi_3$.

The other clauses are easy to check. For instance, for clause 3, we have $\bot \rhd \bot \in U$ by assumption, so $\bot \rhd \bot \in \Phi_3$, as $U \subseteq \Phi_3$. Clauses 1 and 2 are built into Φ_3 by definition. $\Box$

3.4.2 Critical Successors

We now prove some further results towards the proof of The Finite Model Completeness Theorem for ILM 16.

Notation. Given a set of modal formulas Γ, we write $\Gamma \vdash A$ when there is a finite conjunction C of modal formulas in Γ such that $\mathrm{ILM} \vdash C \rightarrow A$.

Definition. We say that Γ is ILM-*consistent* when $\Gamma \nvdash \bot$. Relative to a modal formula A, we say Γ is *maximal* ILM-*consistent* or just *maximal consistent* whenever Γ is ILM-*consistent* and for every subsentence S of A, either $S \in \Gamma$ or $\neg S \in \Gamma$.

Notation. In the following, we consider a fixed finite adequate set Φ, while Γ, Δ will denote maximal ILM-consistent subsets of Φ.

Definition. Δ is a *successor* of Γ, in symbols $\Gamma \prec \Delta$ iff:

1. Both $\Box A, A \in \Delta$ for every $\Box A \in \Gamma$, and

2. There exists a sentence $\Box E$ such that $\Box E \in \Delta$, but $\Box E \notin \Gamma$.

Definition. For a modal formula C, we say that Δ is a C*-critical successor* of Γ iff:

1. $\Gamma \prec \Delta$, and

2. $\Box \neg A, \neg A \in \Delta$ for every $A \rhd C \in \Gamma$, i.e. Δ contains no formulas that "asks for" C.

Note that a successor of a $C-$critical successor is also $C-$critical successor of the original set.

Lemma 19. *The following are theorems of* ILM*:*

1. $(\Box \neg B) \rightarrow (B \rhd C)$*, and*

2. $A \rhd (A \wedge \Box \neg A)$.

Proof. First we prove 2:

$$\Box(\Diamond A \to \Diamond(A \wedge \Box\neg A) \tag{3.2}$$

$$(\Diamond A) \rhd \Diamond(A \wedge \Box\neg A) \tag{3.3}$$

$$(\Diamond A) \rhd (A \wedge \Box\neg A) \tag{3.4}$$

$$(\neg\Box\neg A) \rhd (A \wedge \Box\neg A) \tag{3.5}$$

$$(A \wedge \Box\neg A) \rhd (A \wedge \Box\neg A) \tag{3.6}$$

$$A \rhd (A \wedge \Box\neg A) \tag{3.7}$$

The first equation (3.2) is a theorem of GL, and we have that GL $\subset$ ILM. The second equation follows from the first by applying $\Box(A \to B) \to A \rhd B$ (axiom 4 of ILM). The third equation follows from the transitivity of $\rhd$ and the axiom 8. Rewrite $\Diamond$ from above, and use reflexivity of $\rhd$ to get the fourth and fifth equations. Finally, apply axiom 6 on the fourth and fifth equations to get equation 6.

For 1, notice that GL $\vdash \Box\neg B \to \Box(B \to C)$ (apply distribution of $\Box$ over the propositional calculus tautology $\neg B \to (B \to C)$). Use axiom 4 and tautological reasoning to get ILM $\vdash \Box\neg B \to (B \rhd C)$. □

Lemma 20. *If* $\neg(B \rhd C) \in \Gamma$*, then there exists a* C*-critical successor* Δ *of* Γ *such that* $B \in \Delta$.

Proof. By the second part of the above lemma, $\neg\Box\neg B \in \Gamma$, so since Γ is ILM-consistent (by assumption), then $\Box\neg B \notin \Gamma$. Consider:

$$\Psi = \{D, \Box D | \Box D \in \Gamma\} \cup \{\neg A, \Box\neg A | A \rhd C \in \Gamma\} \cup \{B, \Box\neg B\} \tag{3.8}$$

By the adequacy conditions of Φ, $\Psi \subseteq \Phi$. If we have that Ψ is ILM-consistent, we can take Δ to be the competition of Ψ, that is, a maximal consistent subset of Φ containing Ψ, and we are done. So, assume for contradiction that Ψ is not ILM-consistent. Then $\Gamma \vdash \perp$. We can write this as $D_1, \dots D_k, \Box D_1, \dots \Box D_k \vdash B \wedge \Box\neg B \to A_1 \vee \dots \vee A_m \vee \Diamond(A_1 \vee \dots \vee A_m)$, where $D_i \in \Gamma$ and $A_i \rhd C \in \Gamma$. (we agree here that the empty disjunction is $\perp$ and the empty conjunction is $\top$) So, $\Box D_1, \dots \Box D_k \vdash \Box(B \wedge \Box\neg B \to A_1 \vee \dots \vee A_m \vee \Diamond(A_1 \vee \dots \vee A_m))$, as ILM contains GL, and we can apply necessitation. Then, $\Box D_1, \dots \Box D_k \vdash (B \wedge \Box\neg B) \rhd (A_1 \vee \dots \vee A_m \vee \Diamond(A_1 \vee \dots \vee A_m))$, by application of axiom 4, $\Box(A \to B) \to (A \rhd B)$.

Now, $\Box D_1, \ldots \Box D_k \vdash (B \wedge \Box\neg B) \rhd (A_1 \vee \cdots \vee A_m)$, as $A \vee \Diamond A \rhd A$ is a theorem of ILM, and $\rhd$ is transitive. Then, as the D_i's $\in \Gamma$, and, by the first part of the last lemma, $B \rhd B \vee \Box\neg B$. we have:

$$\Gamma \vdash B \rhd A_1 \vee \cdots \vee A_m. \tag{3.9}$$

Finally, $\Gamma \vdash B \rhd C$, as, for each $i, A_i \rhd C \in \Gamma$. But this contradicts that $\neg(B \rhd C) \in \Gamma$ and the assumed consistency of Γ. So, Ψ is consistent, as desired, and the Lemma holds. □

Lemma 21. *Let $B \rhd C \in \Gamma$. Then, if there exists an E-critical successor Δ of Γ with $B \in \Delta$, then there also exists an E-critical successor Δ' of Γ with $C \in \Delta'$.*

Proof. As above, let:

$$\Psi = \{D, \Box D | \Box D \in \Gamma\} \cup \{\neg F, \Box\neg F | F \rhd E \in \Gamma\} \cup \{C, \Box\neg C\}. \tag{3.10}$$

Then, by adequacy of $\Phi, \Psi \subseteq \Phi$. $\Box\neg C \notin \Gamma$. If it were, applying axiom 7 to the assumption that $B \rhd C \in \Gamma$, we get that $\Diamond B \to \Diamond C \in \Gamma$. Then, $\Box\neg B \in \Gamma$, and so $\neg B \in \Delta$, contradicting our assumption that $B \in \Delta$ and the ILM-consistency of Δ.

Now suppose for contradiction that Ψ is ILM-inconsistent. Similar to (3.9) above, we have $\Gamma \vdash C \rhd F_1 \vee \cdots \vee F_m$, then $\Gamma \vdash B \rhd F_1 \vee \cdots \vee F_m$ (since $B \rhd C$) hence $\Gamma \vdash B \rhd E$. If $m = 0$, then by the second equation $\Gamma \vdash B \rhd \perp$. Then, $\Gamma \vdash \Box\neg B$, and so $\Box\neg B \in \Gamma$. But this contradicts the fact that $B \in \Delta$ and Δ is a consistent successor of Γ. So $m > 0$. As $\Psi \subseteq \Phi$, then E is the consequent of some $\rhd$ formula in Φ. Since $B \rhd C \in \Gamma$, and by assumption $\Gamma \subseteq \Phi$, B is the antecedent of some $\rhd$ formula in Φ. By the adequacy condition, $B \rhd E \in \Phi$. Then, as $\Gamma \vdash B \rhd E$ and Γ is a maximal consistent subset of Φ, we have that $B \rhd E \in \Gamma$. As Δ is assumed to be an E-critical successor of Γ, then $\neg B \in \Delta$. But this contradicts our assumption that Δ is consistent and that $B \in \Delta$. So, Ψ is ILM-consistent. Finally, take Δ' as a maximal consistent subset of Φ that includes Ψ. Then Δ' is an E-critical successor of Γ with $C \in \Delta$, as desired, and the Lemma is proved. □

3.4.3 Proof

We proceed with the proof of the Finite Model Completeness of ILM now that we have the preceding results about adequate sets and critical successors. So assume $\nvdash_{\text{ILM}} A$.

By Lemma 18, we are guaranteed a finite adequate set Φ such that $\neg A \in \Phi$. Furthermore let Γ be a maximal-consistent subset of Φ containing $\neg A$. We will now construct the model $\mathbf{W}$ with root b so that $b \Vdash \neg A$.

Definition. Define the *depth* of a maximally consistent subset Δ of a set Φ as the length of the longest chain of critical successors in Φ that all contain Δ. So if is the longest such chain is $\Delta = \Delta_1 \prec \Delta_2 \prec \cdots \prec \Delta_n$, then the depth of Δ is n

Remember that we have no restrictions whatsoever on the set of an ILM-model. The underlying set W_Γ of $\mathbf{W}$ will be built relative to Γ. Define W_Γ as the set of all pairs $\langle \Delta, \tau \rangle$ such that:

1. The first coordinate Δ is a maximal ILM-consistent subset of Φ that contains Γ in the sense that either $\Gamma \prec \Delta$ or $\Gamma = \Delta$, and

2. The second coordinate τ is a finite sequence of formulas from Φ, such that its length does not exceed the depth of Γ minus the depth of Δ.

Note then that 2. above implies that $\langle \Gamma, \tau \rangle \in W_\Gamma$ iff τ is a sequence of length zero i.e. the empty sequence.

Notation. Given a pair $\langle \Delta, \tau \rangle = w$, denote by $(w)_0$ the first coordinate Δ and by $(w)_1$ the second coordinate τ.

Let the relation R of $\mathbf{W}$ be defined as follows:

$$wRw' \iff (w)_0 \prec (w')_0 \land (w)_1 \subseteq (w')_1. \tag{3.11}$$

Definition. We say that w' is a *C-critical R-successor* of w if the set $(w')_0$ is a C-critical successor of $(w)_0$ and $(w')_1$ is of the form $(w)_1 * \langle C \rangle * \tau$ (where $*$ is the concatenation operator and τ is arbitrary).

Then the relation S_w of $\mathbf{W}$ holds between w' and w'' (write $w'S_w w'$) exactly when:

1. wRw' and wRw'',

2. $(w')_1 \subseteq (w'')_1$,

3. For each A such that $\Box A \in (w')_0$, also $\Box A \in (w'')_0$, and

4. If w' is a C-critical R-successor of w, then so is w".

For the relation $\Vdash$ of $\mathbf{W}$, for p atomic and $w \in W_\Gamma$, define:

$$w \Vdash p \text{ iff } p \in (w)_0. \tag{3.12}$$

Finally, let $\langle \Gamma, \varnothing \rangle = b$ be the root of $\mathbf{W}$.

It is routine to verify that $\mathbf{W} = \langle W_\Gamma, R, S, b \rangle$, with $S = \{ S_w \mid w \in W_\Gamma \}$, as defined above, is a finite ILM-model. To show that $\mathbf{W}$ is our desired counter-model to A (and hence that the Theorem holds), it will be enough to prove the following:

Lemma 22. *For each $A \in \Phi$ and $w \in W_\Gamma$, $w \Vdash A \iff A \in (w)_0$*

This proof is by induction on the complexity of formulas. The cases for the logical connectives $\bot, \rightarrow, \neg$, and $\vee$ are straightforward. For instance for $\bot$, $w \nVdash \bot$ (always), so by equation (3.12) $\bot \notin (w)_0$. The reverse direction is entirely similarly, and so we get the lemma for $\bot$. We can restrict our attention to proving the lemma for $B \rhd C$ granted the lemma holds for both B and C. The reason for this is that ILM $\vdash \Box B \leftrightarrow \neg B \rhd \bot$. So we can transform any $\Box$ formula to one with only $\rhd$'s appearing. Hence if we prove the lemma for $B \rhd C$ then the lemma holds for $\Box$-formulas as well. So to prove the Lemma have to show that:

$$B \rhd C \in (w)_0 \iff w \Vdash B \rhd C, \text{ i.e.} \tag{3.13}$$

$$B \rhd C \in (w)_0 \iff [\forall w'(wRw' \wedge C \in (w')_0 \Rightarrow \exists w''(w' S_w w'' \wedge C \in (w'')_0))] \tag{3.14}$$

($\Leftarrow$) Suppose $B \rhd C \notin (w)_0$. Then $\neg(B \rhd C) \in (w)_0$ by assumption that the the first coordinates of all elements of W_Γ are maximal consistent sets, in particular $(w)_0$. To show that $w \Vdash \neg(B \rhd C)$, we need to show that:

$$\exists w'(wRw' \wedge B \in (w')_0 \wedge \forall w''(w' S_w w'' \Rightarrow \neg C \in (w'')_0)). \tag{3.15}$$

By Lemma (20), $(w)_0$ has C-critical successor $(w')_0$ with $B \in (w')_0$. Let $w' = \langle w'_0, w'_1 \rangle$, where $w'_1 = (w)_1 * \langle C \rangle$. Then $w' \in W_\Gamma$ and moreover w' is a C-critical R-successor of w. Now consider any w'' such that $w' S_w w''$. By (4) of the definition for the relation S_w and the fact that w' is a C-critical R-successor of w, w'' is also C-critical R-successor of w. Then as $C \rhd C \in (w)_0$, we have that $\neg C \in (w'')_0$ by the definition of C-critical successors. This shows equation (3.15), and hence ($\Leftarrow$), as desired.

($\Rightarrow$) Suppose $B \rhd C \in (w)_0$ and that both wRw' and $B \in (w')_0$. Consider the set $\{ \Box D \mid \Box D \in (w')_0 \}$. Because this is a subset of Φ and we have assumed that $B \rhd C \in \Phi$, the last condition of the adequacy of Φ insures that there are sentences $B', C' \in \Phi$ which are ILM-equivalent to $B \wedge \Box D_1 \wedge \cdots \wedge \Box D_n, C \wedge \Box D_1 \wedge \cdots \wedge \Box D_n$ respectively. By adequacy again, $B' \rhd C' \in \Phi$. Then since ILM $\vdash B \rhd C \to B' \rhd C'$ (axiom 9 of ILM) and $(w)_0$ is a maximal consistent subset of Φ (by assumption), we have that $\neg(B' \rhd C') \notin (w)_0$, and hence $B' \rhd C' \in (w)_0$.

Now as $(w)_0 \subseteq (w')_0$ (by assumption), we have that $\Box D_1, \ldots, \Box D_n \in (w')_0$. Also $B \in (w')_0$ (by assumption). Hence $B, \Box D_1, \ldots, \Box D_n \in (w')_0$. Since $(w')_0$ is also a maximal consistent set and $B' \in \Phi$ is ILM-equivalent to $B \wedge \Box D_1 \wedge \cdots \wedge \Box D_n$, reasoning as above we get that $B' \in (w')_0$.

First, assume that w' is an E-critical R-successor of w for some modal formula E. Then $(w')_1 = (w)_1 * \langle E \rangle * \tau$ for some τ by definition 3.4.3. Then by Lemma 21, there is an E-critical w''_0 of $(w)_0$ such that $C \in w''_0$. As w'' is a maximal consistent, $C, \Box D_0, \ldots, \Box D_n \in (w'')_0$ by similar reasoning as above. Now let $w'' = \langle w''_0, w''_1 \rangle$, where $w''_1 = w'_1$. By Definition (3.4.2), as all modal formulas $\Box D$ of $(w')_0$ belong to $(w'')_0$, $(w')_0$ cannot be a successor of $(w'')_0$. Then as $\prec$ is a linear ordering among maximal consistent subsets of Φ (modulo non-boxed formulas), the depth of $(w'')_0$ is no greater than the depth of $(w')_0$. Because $w' \in W_\Gamma$ the length of $(w')_1$ does not exceed the depth of Γ minus the depth of $(w')_0$. Combining these facts, we get that the length of $(w'')_1$ does not exceed the depth of Γ minus the depth of $(w'')_0$. Together with the fact (by construction) that $(w'')_0$ is a maximal consistent subset of Φ, we have from the definition of W_Γ that $(w'') \in W_\Gamma$. It is easy to verify then that $w' S_w w''$ and hence equation (3.14) holds.

Assume on the other hand that w' was not an E-critical R-successor. Then all we know from the assumptions is that $(w)_0 \prec (w')_0$. But every successor is also a $\bot$-critical successor. So we can still apply Lemma 21 and proceed just as above to verify equation (3.14) to conclude the proof of Lemma 22. This finishes the proof of Lemma 22.

To finish the proof of the Finite Model Completeness Theorem, and hence the Main Theorem 16 of this chapter, recall that we defined the root of $\mathbf{W}$ as $b = \langle \Gamma, \varnothing \rangle$. By the stipulation of Γ at the beginning of the Theorem, $\neg A \in \Gamma$. So by Lemma 22, $b \Vdash \neg A$. We have from above that $\mathbf{W}$ is a finite ILM-model. So under the assumption that ILM $\nvdash A$, we have a finite ILM-model with root b such that $b \models \neg A$. Note that as ILM is both sound and complete with respect to a finite class of models, we also have the decidability of theory ILM with this result. $\Box$

3.5 Simplified Model Completeness

ILM-models are defined in Section 3.3, while simplified ILM-models will be defined below. Simplified models are simplified in the sense that they have a single S relation instead of the set $\{ S_w \mid w \in W \}$ in ILM-models.

In this section, we prove that for every finite ILM-model, there is a simplified ILM-model such that the two satisfy exactly the same modal formulas at their respective roots. Applying the Model Soundness and Completeness Theorem for ILM 16 will then give the Simplified Model Soundness and Completeness Theorem for ILM 23 with respect to simplified ILM-models. This will be used extensively in the proof of the Main Result of this paper, presented in Chapter 4. The original result is proved by Albert Visser in [12], while the presentation below follows [1].

3.5.1 Simplified ILM-Models

Definition. A *simplified ILM-frame* $\langle W, R, S\rangle$ is a GL-frame $\langle W, R\rangle$ together with just one additional relation S between nodes that satisfies:

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

Definition. A *simplified* ILM-*model* is given by a simplified ILM-frame together with (as in Definition (3.3)) a forcing relation $\Vdash$ which commutes with the boolean connectives and obeys:

$$x \Vdash \Box A \iff \forall x(xRy \Rightarrow y \Vdash A),$$

$$x \Vdash A \rhd B \iff \forall y(xRy \wedge y \models A \Rightarrow \exists z(xRz \wedge ySz \wedge z \Vdash B)).$$

Notation. $\mathcal{K}_{SM}$ will denote the class of simplified ILM-frames.

Simplified Model Soundness and Completeness Theorems for ILM 23. *For each A, $\vdash_{\text{ILM}} A$ if and only if $\mathcal{K}_{SM} \models A$.*

The $(\Leftarrow)$ direction is routine to verify and amounts to showing that all the axioms of ILM are valid at every node of every simplified ILM-model and that each rule of inference preserves validity in every simplified ILM-model.

The $(\Rightarrow)$ is shown by applying the Bisimulation Theorem 24 (below) to the model $\mathbf{W}$ provided by the Model Completeness Theorem for ILM 16 to obtain the desired simplified ILM-model $\mathbf{W}'$.

3.5.2 Bisimulation

Bisimulation Theorem 24. *For every* ILM*-model* $\mathbf{W}$ *with root* b *there is a simplified* ILM*-model* $\mathbf{W}'$ *with root* b' *such that for all modal formulas* A*,* $b \Vdash A \iff b' \Vdash A$*.*

We will explicitly build the models $\mathbf{W}, \mathbf{W}'$ given models $\mathbf{W}', \mathbf{W}$ respectively.

Building an ILM-model $\mathbf{W}'$ from a simplified ILM-model $\mathbf{W} = \langle W, R, S, b, \Vdash \ \rangle$ is easy: We take the same underlying set W and relation R as $\mathbf{W}$. Build the relation S_w for each $w \in W$ by setting:

$$wRw' \wedge wRw'' \wedge w'Sw'' \Rightarrow w'S_w w'' \tag{3.16}$$

The forcing relation $\Vdash'$ for $\mathbf{W}$ is set to agree on all nodes with the old forcing relation $\Vdash$ for all atomic modal formulas. It is easy to check then that the two relations will agree on all modal formulas.

We can see the new ILM-model $\mathbf{W} = \langle W, R, \{S_w\}, \Vdash \rangle$ as just the old model $\mathbf{W}'$ with the addition of the relations $\{S_w\}$ built according to equation (3.16). This construction of $\mathbf{W}$ from $\mathbf{W}'$ will be called the *induced* ILM*-model.* As mentioned above a simplified ILM-model and its induced ILM-model will agree on all modal formulas on all nodes, in particular their roots. We can then think of the class of simplified ILM-models as a subset of the class of ILM-models in that each simplified ILM-models has a naturally induced ILM-model.

Building a simplified ILM-model from an ILM-model on the other hand is a more involved matter. We must "collapse" the set of relations $\{S_w\}$ in $\mathbf{W}'$ into a single relation S in $\mathbf{W}$. Although the two will agree on all formulas on their roots, the two will in general not share the same underlying set. The relationship between $\mathbf{W}$ and the induced model of $\mathbf{W}'$ will be called 'bisimulation'.

Definition. Let $\mathbf{W}$ and $\mathbf{W}''$ be two ILM-models. Let $w\beta w''$ be a relation between elements of W and W''. Then $\mathbf{W}$ and $\mathbf{W}''$ *bisimulate* each other (or β is a *bisimulation* between the two) if and only if the following hold:

1. $b\beta b''$, where b and b'' are the respective roots of W and W'',

2. $x\beta x'' \rightarrow (x \Vdash A \leftrightarrow x'' \Vdash'' A)$, for A atomic,

3. For all $x, y, z \in W$, $x'', y'', z'' \in W''$ with $x\beta x''$, the following holds:
   $\forall y[xRy \to \exists y''(y\beta y'' \wedge x''R''y'' \wedge \forall z''(y''R_{x''}z'' \to \exists z(z\beta z'' \wedge yS_xz)))]$,
4. Conversely for all $x, y, z \in W$, $x'', y'', z'' \in W''$ with $x''\beta x$:
   $\forall y''[x''Ry'' \to \exists y(y''\beta y \wedge xR''y \wedge \forall z(yR_xz \to \exists z''(z''\beta z \wedge y''S''_{x''}z'')))]$,

We can check that β is reflexive, symmetric, and transitive, and hence is an equivalence relation on the class of ILM-models. Furthermore:

Lemma 25. *If β is a bisimulation between $\mathbf{W}$ and $\mathbf{W}''$ and $x\beta x''$, then for all modal formulas, $x \Vdash A \leftrightarrow x'' \Vdash A''$.*

Proof. Proceed by induction on complexity of formulas. The cases for $\bot$ and the logical connectives are routine and left for the reader. As before, we will use that ILM $\vdash \Box A \leftrightarrow \neg\Diamond\neg A$ and ILM $\vdash \Box A \leftrightarrow \neg A \rhd \bot$ to proceed to the case where $A = B \rhd C$ and the Lemma holds for B and C. By symmetry of the definition of β, we just need to show one direction of the Lemma. So assume $x \nVdash A$. By Definition 3.3, there is a y with $xRy, y \Vdash B$, and for all z such that yS_xz, we have $z \nVdash C$. As $x\beta x''$ by assumption of the Lemma, there is a y'' guaranteed by part 4 of the definition of β such that:

$$y''\beta y \wedge xR''y \wedge \forall z(yR_xz \to \exists z''(z''\beta z \wedge y''S_{x''}z'')) \tag{3.17}$$

From the induction hypothesis as we have $y\beta y''$, then also $y'' \Vdash'' B$. The claim now will be that y'' witnesses that $x'' \nVdash C$. For a contradiction assume not, so that there is a z'' such that both $y''S''_{x''}z''$ and $z'' \Vdash C$. Then by equation (3.17) we would have a z such that both $z\beta z''$ and yS_xz. But then by the induction hypothesis, we would have that $z \Vdash \neg C$, contrary to above that $z \Vdash C$. So z'' does not exist, and y'' witnesses that $x'' \nVdash A$, as desired. $\Box$

We return to the proof of the Bisimulation Theorem 24 and show that $\mathbf{W}$ and $\mathbf{W}'$ (defined below) will bisimulate each other. Define $\mathbf{W}' = \langle W', R', \{S'_{w'}\}, \Vdash' \rangle$ relative to $\mathbf{W}$ as follows:

1. W' is the set of all finite sequences $\langle w_1, \ldots, w_n\rangle$ with elements w_i from W such that both $x_1 = b$ and for every $0 < i < n$, either x_iRx_{i+1} or $x_iS_jx_{i+1}$ with $j < i$.
2. $b' = \langle b\rangle$

3. For for all atomic p, $\langle x_1, \dots x_n \rangle \Vdash' p$ iff $x_n \Vdash p$.

4. $\langle x_1, \dots, x_m \rangle S' \langle y_1, \dots, x_n \rangle$ iff $\forall i \leq m (x_i = y_i)$ and $m \leq n$. That is, S' is the relation of (not necessarily proper) end-extensions among sequences.

5. $\langle x_1, \dots, x_m \rangle R' \langle y_1, \dots, x_n \rangle$ iff $\langle x_1, \dots, x_m \rangle S' \langle y_1, \dots, x_n \rangle$ with $m < n$, $\exists k : m \leq k < n$ so that $x_k R x_{k+1}$, and $\forall j \exists s : m \leq s < j < n$ so that $y_j S_{y_s} y_{j+1}$.

Note in 5. above there will be a maximum $k : m \leq k < n$. Without loss of generality we can consider this k and assume without loss of generality that $\forall j > k \neg (x_j R x_{j+1})$.

Lemma 26. *If $\langle x_1, \dots, x_m \rangle R' \langle y_1, \dots, y_n \rangle$, then $x_m R y_n$.*

Proof. By 5. above we have that $n > m$. Also either $y_{n-1} R y_n$ or $\exists s : m \leq s < n$ so that $y_{n-1} S_{y_s} y_n$ and hence by the definition of S_w that $y_s R y_n$. So in either case $\exists j : m \leq j < n$ so that $y_j R y_n$. By 1. above, we have for $\vec{y}$ that or every $i < n$, either $x_i R x_{i+1}$ or $x_i S_j x_{i+1}$ with $j < i$ Then iteratively apply the property that $\forall z : w S_z w' R w'' \to w R w''$ to lower j to m to get $x_m R y_n$. □

Now it is routine to verify that $\mathbf{W}'$ is a simplified ILM-model as defined in Definition 3.5.1. In particular we use Lemma 5 to conclude that R' is also converse well-founded from the fact that R is converse well-founded.

Lemma 27. *If $\langle \dots x \rangle R' \langle \dots x, y \rangle S' \langle \dots, x, y, \dots, z \rangle$ and $\langle \dots x \rangle R' \langle \dots x, y, \dots, z \rangle$, then $y S_x z$ (possibly $y = z$).*

Proof. As $\langle \dots x \rangle R' \langle \dots x, y \rangle$, from 5. above, we have that $k = x$ and hence $x R y$. Since $\langle \dots x \rangle R' \langle \dots x, y, \dots, z \rangle$ by the same definition there is an R-step between adjacent elements somewhere between x and z. There are two cases to consider:

1. The R-step happens between x and y. If S_x steps are taken between every node and its immediate successor, then because S_x is defined to be transitive, we have $y S_x z$ as desired. If not, then there are u, v, w, $u \neq x$ such that $v S_u w$. Then from this we have $u R w$. Without loss of generality we can assume that the last non-S_x step is taken between v and w and rewrite $\vec{z}$ as:

$$\langle \dots, z \rangle = \langle \dots, x, y, \dots, u, \dots, v, w, \dots, z \rangle \tag{3.18}$$

with possibly $y = u$ and $w = z$. From the definition of elements of W' we can conclude that we can go between any two elements in $\vec{z}$ (in particular y and u) with S_i moves of (not necessarily identical) indices. (Note this point is not immediate as this definition allows for an R move between adjacent elements of the sequence $\vec{z}$. But if $x_k R x_{k+1}$, we can examine cases and conclude that also $x_k S_i x_{k+1}$ for some $i < k$.) Using uRw and $\forall z : wS_z w' R w'' \rightarrow wRw''$ again we conclude that yRw. Then from this and the assumption xRy we get that xRy, xRw, and yRw, i.e. that $yS_x w$. But we assumed that only S_x steps were taken after w. So by the transitivity of any S_i we get that $yS_x z$, as desired.

2. The R-step happens somewhere between y and z. In this case cRd for some c, d such that:

$$\langle \ldots, z \rangle = \langle \ldots, x, y, \ldots, c, d, \ldots, z \rangle \tag{3.19}$$

with possibly $y = c$ and $d = z$. We know that between d and y S_i steps occur with $x \leq i \leq y$. As above, using cRd and $\forall z : wS_z w' R w'' \rightarrow wRw''$ we conclude that xRd, yRd. We already have that xRy. Therefore $yS_x d$. Now reason with d as we did with y in the proof above to get $dS_x z$. Combine this with $yS_x d$ and that every S_i is transitive to get $yS_x z$, as desired. This completes the proof of the Lemma. □

Now let $\mathbf{W}''$ be the ILM-model induced by the simplified ILM-model $\mathbf{W}'$ through $wRw' \wedge wRw'' \wedge w'Sw'' \Rightarrow w'S_w w''$ (equation (3.16)). Remember $\mathbf{W}'$ and $\mathbf{W}''$ share the same underlying set W, relation R, and forcing relation $\Vdash$.

Lemma 28. *Define $w\beta w''$ iff $x_n \beta \langle x_1, \ldots, x_n \rangle$. Then $\mathbf{W}$ and $\mathbf{W}''$ bisimulate each other through β.*

Proof. Clauses 1 and 2 of Definition 3.5.2 saying that the two models share the same root and that their forcing relations agree on atomic p are true by construction of $\mathbf{W}''$ relative to $\mathbf{W}$.

For clause 3, suppose $x\beta x''$ and xRy, i.e. $x\beta\langle \ldots, x\rangle$ and xRy. Then to find a y'' such that $y\beta y''$ and $x''R''y''$ take $y'' = \langle \ldots, x, y\rangle$, which satisfies $y\beta\langle \ldots, y\rangle$ and $\langle \ldots, x\rangle R'' \langle \ldots, x, y\rangle$. Now consider any z'' such that $y'' R_{x''} z''$. Rewriting this for the model $\mathbf{W}''$, consider any $\langle \ldots, x, y, \ldots, z\rangle$ such that $\langle \ldots, x\rangle R'' \langle \ldots, x, y\rangle$, $\langle \ldots, x\rangle R'' \langle \ldots, x, y, \ldots, z\rangle$, and $\langle \ldots, x, y\rangle S''' \langle \ldots, x, y, \ldots, z\rangle$

all hold. Then $\langle \dots x\rangle R''\langle \dots x, y\rangle S''\langle \dots, x, y, \dots, z\rangle$ and $\langle \dots x\rangle R''\langle \dots x, y, \dots, z\rangle$ both hold. By Lemma 27, yS_xz, as desired.

For clause 4, suppose $x\beta\langle \dots, x\rangle$ and $\langle \dots, x\rangle R''\langle \dots, x, \dots y\rangle$. By Lemma 26, xRy. Suppose yS_xz. We want to have a z'' such that $z\beta z''$ and $y''S''_{x''}z''$. I claim $z = \langle \dots, x, \dots, y, z\rangle$ will work. That $z\beta z''$ holds is clear. Rewriting $y''S''_{x''}z''$ we need to show that $\langle \dots, x\rangle R''\langle \dots, x, \dots y\rangle$, $\langle \dots, x\rangle R''\langle \dots, x, \dots y, z\rangle$, and $\langle \dots, x, y\rangle S''\langle \dots, x, y, \dots, z\rangle$ all hold. The last is clear from the definition of S'' as the relation of end-extensions. We have the first by assumption, which together with the assumption that yS_xz (and hence yRz) implies the second. This proves the Lemma.

To finish the proof of the Bisimulation Theorem 24, take $\mathbf{W}''$ as define above. As both the bisimulation and the the inducement maps preserve which formulas are satisfied at the roots of the models they act on, we have what we desire.

□

Chapter 4

Main Result

In this section, we will show that the theorems of ILM exactly characterize the universally valid inferences one can make concerning interpretability over the base theory PA. Taking PA as a base theory means that the relation we will be analyzing will be between two formulas written in the language of PA. The relation will hold if and only if the finite extension of PA through addition of A interprets the finite extension of PA through addition of B. So what we want is that the modal formula $A \rhd B$ represents the statement: PA proves that 'PA $+ A$ interprets PA $+ B$'. To make this idea explicit and prove it, we will introduce the notion of an arithmetical realization. This will be a map that 'realizes' modal formulas as PA formulas. Importantly, realizations will respect connectives, map propositional variables to PA sentences, and map the symbol $\rhd$ to a formalization (in PA) of the interpretability relation over PA. Then to rephrase the purpose of this section, we will show that ILM proves a statement if and only if PA proves all arithmetical realizations of the statement.

As a simple motivating example, $\vdash_{\mathrm{ILM}} A \rhd B \wedge B \rhd C \to A \rhd C$. Then with the Main Result we can easily conclude that the interpretability relation of single-sentences extensions of PA is transitive. The advantage of such a general approach is that there is no need to consider any specific details of the sentences or the interpretations that witness the assumptions.

4.1 Arithmetical Realizations

Definition. By an *arithmetical interpretation*, we mean a mapping $\star$ from the modal language of ILM to the language of PA which commutes with the propositional connectives and such that $(A \rhd B)^\star = \mathrm{Interp}_{\mathrm{PA}}(\ulcorner A \urcorner, \ulcorner B \urcorner)$, where $\mathrm{Interp}_{\mathrm{PA}}(\ulcorner A \urcorner, \ulcorner B \urcorner)$ is a formalization (in PA) of 'PA + $A^\star$ interprets PA + $B^\star$'.

We are now in a position to state the Main Result of this paper. Alessandro Berarducci in [1] and Volodya Shavrukov in [11] independently obtained the original proofs of this result. This presentation follows the former paper.

Main Result. *(Outside* PA*) For all arithmetical realizations* $\star$*,*

$$\mathrm{ILM} \vdash A \textit{ if and only if } \mathrm{PA} \vdash A^\star \tag{4.1}$$

The ($\Rightarrow$) direction is called the *Arithmetical Soundness of* ILM. This direction is much simpler than the reverse. As with other soundness proofs, it is enough to show that all realizations of the axioms of ILM are provable in PA, and that the rules of inference preserve provability. For instance, any realization of Axiom 8 is the statement that PA + Con(PA + A) interprets PA + A. This all these realizations are provable follows from the Formalized Gödel's Completeness Theorem.

The ($\Leftarrow$) direction is called the *Arithmetical Completeness of* ILM. It involves more complicated reasoning. The key element will be a systematic procedure that takes an arbitrary unprovable formula of ILM, acts on its counter-model guaranteed by the Simplified Model Completeness Theorem in the previous chapter, and constructs an induced arithmetical realization. Moreover the induced realization of the original ILM-unprovable formula will be unprovable in PA. For example, if the ILM-unprovable formula is of the form $A \rhd B$, then the procedure will generate a PA-unprovable statement that is the formalization of the statement: 'PA + A does not interpret PA + B'. Of course, the constructed statement will be in the language of PA, so it will ostensibly be a statement about the natural numbers. But by a numbering of the vocabulary of PA, the number-theoretic statement will mirror the syntactical relation that holds only between extensions of PA that do not interpret one another. The part below will follow [1].

4.2 Arithmetical Completeness

Proof. Assume $\nvdash_{\mathrm{ILM}} A$. By the Simplified Model Completeness Theorem, there is a simplified ILM-model, call it $\mathbf{W} = \langle W, R, S, b, \Vdash\rangle$, such that $b \Vdash \neg A$.

Without loss of generality assume that in the model $b = 1$ and $0 \notin W$. We will now slightly modify our model for purposes of the proof. We will adjoin a *new 0 root* to the model $\mathbf{W}$. The motivation for this is that we are going to define a function F that defines a process that moves along the model with the new 0 root. We will eventually want to show that in the standard model that the process never leaves the 0 root. But we will also show that if a node is R-connected to the 0 root, that the following statement will hold in PA: 'PA plus 'the limit is 1' is consistent'. But if this is consistent, then PA $\nvdash$ 'the limit is not 1'. We will build the induced interpretation $*$ to use this fact together with the assumption that $1 \Vdash \neg A$ to obtain PA $\nvdash A^*$, as detailed below.

With this in mind, extend R and S by setting $0Rx$ for all x in W, and $0Sx$ for all x in $W \cup \{0\}$. Also extend the forcing relation so that $0 \Vdash \ A \iff 1 \Vdash \ A$, for all A atomic. This gives a new simplified ILM-model with underlying set $W \cup \{0\}$ and root 0. The two models agree on their common domain in the sense that for all formulas A, if $x \Vdash A$ in one model, then $x \Vdash A$ in the other. From this point on, we will denote by $\mathbf{W}$ the model with the new 0 root.

4.2.1 Definition of F

We will think of F as a process that starts at the 0 root and moves along the model's frame in discrete, numbered steps. To define F, we will first need to introduce the concept of the rank of a node at a stage:

Definition. Let $x \in W$. We define *rank(x,n)*, the *rank of x at stage n*, as the smallest number $i \leq n$ such that PA$_i$ proves that $L \neq x$ with a proof of Gödel number $\leq n$. If i does not exists, then we define $rank(x, n)$ to be the ordinal ω. Note that if the rank of an element y at stage n is less than ω, then the rank of y will be $\leq n$. This holds as n bounds the size of proofs of $L = y$, exactly what we defined rank to be. Intuitively, the smaller $rank(y, n)$ is, the more inconsistent is the fact that $L = y$.

We will also need a (provably) infinitely repeating primitive recursive coding of the nodes of $W \cup \{0\}$. This will be a function from $\mathbb{N}$ to $\mathbb{N}$, or from codings to nodes, such that each node has countably infinite many inverse elements (codings). For example, we can set

$$\text{‘}n \text{ codes } x\text{’ iff ‘}(\exists y \leq n)\ (n = 2^y(2x+1))\text{’}. \tag{4.2}$$

Definition. Define in PA the *function* F as follows: Set $F(0) = 0$. Assume that $F(n) = x$, and F has been defined for every $m \leq n$. Define $F(n+1)$ as follows:

1. Suppose that n codes an element y in $W \cup \{0\}$ and that xRy. If $rank(y, n) \leq \omega$, define $F(n+1) = y$.
2. Suppose that n codes an element y in $W \cup \{0\}$, $\neg xRy$, and xSy. Suppose further that $rank(y, n) < rank(x, n)$. Then $rank(y, n) < \omega$, and from the fact above, $rank(y, n) < n$. But F has already been defined for all values up to n. So $rank(y, n) = a$, for some a. If aRy, then define $F(n+1) = y$.
3. If neither of these hold, define $F(n+1) = 1$.

Definition. Define in PA the constant L as:

1. If F has a limit, define $L =$ the limit of F,
2. Otherwise, define $L = 0$.

Despite the apparent circularity of the definition of the function F and its limit L, we can nonetheless obtain a function which satisfies all these conditions through the Diagonal Lemma 3.

4.2.2 Properties of the Function F

Notation. Boldfaced notation for numerals in PA will be omitted where the meaning is clear from context. For instance, we will write $\text{Con}(\text{PA} + (L = x))$ in place of $\text{Con}(\text{PA} + (L = \mathbf{x}))$.

Theorem 29. *F is a primitive recursive function and L is a definable constant of* PA *such that* PA *proves the formalizations of the following statements:*

($\neg$S$_1$) *For all m, n, if $m \leq n$, then $F(m)SF(n)$,*

($\neg$S$_2$) *L is the limit of the function $F : \omega \to W \cup \{0\}$,*

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

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

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

Proof. We can verify that the function F is primitive recursive based on its definition.

Now work in ACA$_0$ through the proof.

($\neg$S$_1$) According to which of the three clauses of F is satisfied, exactly one the following holds: $F(n)RF(n+1)$, $F(n)SF(n+1)$ or $F(n) = F(n+1)$. In all cases we will have $F(n)SF(n+1)$, as both S is reflexive and $R \subseteq S$ by the definition of S in ILM-models. That the claim holds follows from repeated applications of this fact.

($\neg$S$_2$) By the definition of simplified ILM-models, the relation R is converse well-founded, which means that there is some bound k for the frame such that every R-chain has length $\leq k$. Furthermore because of the relation $xSyRz \Rightarrow xRz$, any set of R-jumps interspersed with S jumps (e.g. $x_1Rx_2Sx_3Rx_4$) will be joined into a chain of consecutive R-jumps (e.g. $x_1Rx_2Rx_4$). Therefore F can only make at most k R-jumps, whether consecutive or not. So eventually F will only be making S-jumps. But by the definition of F, an S-move from x to y will only take place only when the rank of y (at all stages $> n$) is smaller than the rank of x (at stage n). Then if F did not have a limit, we would have an infinitely descending, definable sequence of ranks (i.e. integers). Absurd. In other words the process defined by the function F peters out eventually as both R and S-jumps must come to an end.

(R) From the fact R is converse well-founded, it is an easy consequence that R is not reflexive. So if xRy, then $x \neq y$. For a contradiction, assume that xRy, $L = x$, and that $\mathrm{PA} + L = y$ is not consistent. Let n be such that n codes y and n is so large that:

1. F has already reached its limit L at stage n, and

2. There is a proof of $L \neq y$ from PA$_n$ with Gödel number less than n (thus $rank(y, n) < \omega$).

That some number satisfies 1. follows from the assumption that $L = x$. That some number satisfies 2. follows from the assumption that $\mathrm{PA} + L = y$ in not consistent. For both these properties, if n satisfies the property, then all numbers greater than n also do. Then together with the fact that our code (4.2) is infinitely repeating, we have a single n that jointly satisfy the desired properties. But what we have described are exactly the conditions needed for an R-jump from x to y according to the definition of F. Then the limit $L \neq x$, contrary to our assumption that $L = x$. End of contradiction. So in fact $\mathrm{PA} + L = y$ is consistent, as desired.

($\neg$R) By assumption, $L = x$. Now consider any k such that $x = F(k)$. From the definition of F and the fact that F moved to node x (as $L = x$), we conclude that $rank(x, k) < \omega$, and hence $rank(x, k) \leq k$. Because we are working in ACA_0, we can employ model-theoretic notions. So for a contradiction, assume that there is a model $\mathcal{Y}$ of PA and an element y of $\mathcal{Y}$ such that:

$$\mathcal{Y} \models L = y \wedge \neg xRy. \tag{4.3}$$

As $x = F(k)$ is a Σ_1-sentence and the fact that $x = F(k)$, we have $\mathcal{Y} \models$ '$x = F(k)'$. Then we can conclude that k is a standard element of $\mathcal{Y}$. By ($\neg\mathrm{S}_1$), we have that $\mathcal{Y} \models xSy$.

Now, in $\mathcal{Y}$ consider the last step taken by F before reaching y. In other words, consider the n, w such that $F(m) = w$, $w \neq y$, and $\forall n > m(F(n) > y)$. Since by assumption $L = x$, then n (chosen as the penultimate step before the limit in the model $\mathcal{Y}$) must be a nonstandard element of the model $\mathcal{Y}$. As $\mathcal{Y} \models n > k$, $F(k) = x$, and $F(n) = w$, by ($\neg\mathrm{S}_1$) we have $\mathcal{Y} \models xSwSy$. Also wRy is impossible, as it would imply (by $xSwRy \Rightarrow xRy$) that xRy, and hence $\mathcal{Y} \models xRy$, contrary to our assumption that $\mathcal{Y} \models \neg xRy$. So $\neg wRy$.

Now let $a = F(rank(y, n))$. As $\neg wRy$, then by the definition of F and the assumption that $L = y$, we must have that F moved from w to y via an S-move. Hence we have that $\mathcal{Y} \models aRy$. Now consider $rank(y, n)$. If we had $\mathcal{Y} \models k \leq rank(y, n)$, then by ($\neg\mathrm{S}_1$) we would have that $\mathcal{Y} \models xSaRy$, and hence $\mathcal{Y} \models xRy$ (by $xSwRy \Rightarrow xRy$ again). But this a contradiction to our assumption that $\mathcal{Y} \models \neg xRy$.

If, on the other hand, $\mathcal{Y} \models k > rank(y, n)$, then $\mathcal{Y} \models \neg\mathrm{Con}(\mathrm{PA}_k{+}L = y)$. This holds because we defined $rank(y, n) = k$ to mean that there is a proof in PA_k of $L \neq y$ of length $< n$. But this again is a contradiction; because we picked k to be standard, the Reflection Theorem 5 guarantees that PA proves the consistency of every finite fragment of itself. End of contradiction. So given the assumption of the claim, we now have (stated outside PA) that: $\neg\mathrm{Con}(\mathrm{PA} + \exists y : L = y \wedge \neg xRy)$, as desired. Note that our original k defined as any stage that $y = F(k)$ became the size of PA_k such that $\mathcal{Y} \models \neg\mathrm{Con}(\mathrm{PA}_k + L = y)$.

(S) Assume that $L = x$. Fix a standard natural k. For a contradiction, assume that PA does not prove the consequent of (S). Then there is a model $\mathcal{M}$ of PA with two elements y, z such that $\mathcal{M} \models L = y \wedge xRz \wedge ySz \wedge \neg\mathrm{Con}(\mathrm{PA}_k + (L = z))$. But then $\mathcal{M}$ sees that $rank(z, k) \leq k$, and hence for sufficiently large m, $rank(z, m) \leq k$. (In other words this last equation states that there is a proof of $L \neq y$ with size less than k. This follows from the last conjunct above and the fact that $rank(z, m)$ is non-increasing for fixed m). By reflection, the fact k is standard, and that $\mathcal{M}$ models $\mathrm{PA} + L = y$, we have that $\mathcal{M}$ is a model of $PA_k + L = y$. Hence, $\mathcal{M} \models \forall m(rank(y, m) > k)$. To reach a contradiction, we will now show that F moves to z, contradicting the fact that $\mathcal{M} \models L = y$. Let n be large enough so that F has reached its limit y at stage n, n codes z, and $rank(z, n) \leq rank(y, n)$. We will want to show that $\mathcal{M} \models F(n + 1) = z$. Let $r \in \mathcal{M}$ be such that $\mathcal{M} \models rank(y, n) = r$. We can compute, outside of the model, the (not necessarily unique) node a that the function F reaches at step r, i.e. an a such $F(r) = a$. Since $L = x$, from property (¬S), we have that aSx. By the absoluteness of Σ_1 statements, we have that $\mathcal{M} \models aSx$. From a formalized version of the ILM-frame property $aSxRz \implies aRz$, we have that that $\mathcal{M} \models aRz$. But now all the conditions are satisfied for F to make an S move from y to z at stage n (cf. part 2. of the definition of F). Then $\mathcal{M} \models F(n+1) = z$, which is the desired contradiction So, in PA, under the assumptions above, we have that for all k, PA proves that for all $y, z \in W \cup \{0\}$, if xRz and ySz, then $\mathrm{Con}(\mathrm{PA}_k + (L = z))$.

This ends the proof of Theorem 29. □

Remark 30. What we have just shown in (S) can be alternatively written as:

$$\begin{aligned} &\mathrm{ACA}_0 \vdash \forall x(x \in \mathbf{W} \wedge x = \lim(f) \\ &\to \forall k(\ulcorner \forall y, z, (y, z \in \mathbf{W} \wedge y = \lim(f) = y \wedge xSz \wedge yRz \\ &\to \neg\mathrm{Bew}_k(\ulcorner z \neq \lim(f) \urcorner)\urcorner \in \mathrm{PA})) \end{aligned} \tag{4.4}$$

Remark 31. (Outside PA) Since $\forall x \in W(\neg xRx)$, property $(\neg R)$ implies that

$$\forall x \in W(\mathrm{PA} + L = x \vdash \neg\mathrm{Con}(\mathrm{PA} + L = x)) \tag{4.5}$$

Then in the intended model ω of number theory, the limit L cannot be any node in W. Combining this with $(\neg S)$ that F is a function on $W \cup \{0\}$, we get that in the standard model the limit L is 0.

By property (R) and by how we adjoined the node 0 to our model so that $0Ry$ for all y in W, we have that the standard model satisfies, for all y in W

$$\mathrm{Con}(\mathrm{PA} + L = y) \tag{4.6}$$

We have that if a particular instance of formula (4.6) is satisfied in the standard model for an element y, then there is a model of $\mathrm{PA} + L = y$. But we have formula (4.6) for each $y \in W$. So for each $y \in W$, $\mathrm{PA} + L = y$ has a model, and hence:

$$\text{for each } y \in W,\ \mathrm{PA} + L = y \text{ is consistent with PA.} \tag{4.7}$$

4.2.3 The Induced Realization

We will now define the *induced arithmetical realization* $*$ such that if $1 \nVdash A$, then $\mathrm{PA} \nvdash A^*$ The realization $*$ will be defined relative to the constant L, which in turn depends on the model $\mathbf{W}$.

Definition. Define the *induced arithmetical realization* $*$ so that it is an arithmetical interpretation and for atomic A, A^* is the sentence of PA is a formalization of the following:

$$\text{`}\exists x \in W \cup \{0\} : L = x \wedge x \Vdash A\text{'}$$

Remember that $*$ will commute with propositional connectives and that $(A \rhd B)^* = \mathrm{Interp}_{\mathrm{PA}}(\ulcorner A \urcorner, \ulcorner B \urcorner)$.

4.2.4 Properties of F Imply the Main Result

Definition. Let C be a formula in $\mathcal{L}(\rhd)$. We say that the realization $\star$ is *faithful* on C iff, in PA, for all $x \in W$, we have:

1. if $x \models C$ and $L = x$, then $C^\star$, and

2. if $x \not\models C$ and $L = x$, then $\neg C^\star$.

Call the first condition *faithful 1* and the second condition *faithful 2.*

Lemma 32. *For all C in $\mathcal{L}(\rhd)$,* PA *proves that the induced interpretation $*$ as defined in Definition 4.2.3 is faithful on C.*

Proof. Work in PA. We will show that each way of building a formula C preserves faithfulness (i.e. a proof by induction on the complexity of C). The interpretation $*$ is faithful 1 on $\bot$ vacuously and faithful 2 by the fact that for all $x, x \not\models \bot$. It is also faithful on C atomic (which closely follows from the definition of $*$). Next, $*$ is faithful on boolean combinations of formulas $*$ is already faithful on. For example, $*$ is faithful on $A \to B$ if $*$ is faithful on both A and B. This follows from observing that $*$ was defined to distribute over boolean combinations. As ILM $\vdash \Box A \leftrightarrow \neg A \rhd \bot$, we have: $*$ is faithful on $\Box A$ assuming it is faithful on A if and only if $*$ is faithful on $A \rhd B$ assuming it is faithful on both A and B. So we only need to check the inductive case $A \rhd B$.

Proposition 33. (PA) *If $x \models A \rhd B$ and $L = x$, then C^*.*

Work in ACA_0 to formalize model-theoretic notions. Suppose that $x \in W$, $L = x$, and $x \models A \rhd B$. We must prove that $(A \rhd B)^*$ which is shorthand for 'PA+A^* interprets PA+B^*', with A^* and B^* as defined in Definition 4.2.3. Assume for contradiction this is not the case. Then by Theorem 12, there is a model $\mathcal{Y}$ of PA + A^* that has no end-extension $\mathcal{Z}$ that models PA + B^*.

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

Take y to be the unique element such that $\mathcal{Y} \models L = y$. As $x \in W$, $L = x$, and $\mathcal{Y} \models L = y$, by property ($\neg$R), we have that xRy. Now use the induction hypothesis: PA $\vdash$ 'I is faithful on A'. Note that this induction hypothesis was assumed outside of PA, yet we wish to use it inside of PA. This issue can be taken care of seeing that this is a Σ_1-assertion, namely the fact that something is provable in PA, and so will hold inside of PA as well. As $\mathcal{Y} \models$ PA, consequently $\mathcal{Y} \models$ 'I is faithful on A'. Hence using the contrapositive of the second part of faithfulness, $\mathcal{Y} \models A^*$ (by construction), and $\mathcal{Y} \models L = y$ (by above), we have that $\mathcal{Y} \models (y \models A)$. This completes the claim that there is an element $y \in \mathcal{Y}$ such that $\mathcal{Y} \models xRy \wedge y \models A$.

Now, in ACA_0, we are assuming $x \models A \rhd B$. But this is Σ_1, so it must hold in all models of PA. In particular it holds in $\mathcal{Y}$, so $\mathcal{Y} \models A \rhd B$. We can use the requirement on ILM-frames imposed by $\rhd$ inside $\mathcal{Y}$, i.e. that $A \rhd B \iff (\forall y)(y \models A \wedge xRy \implies \exists z(xRz \wedge ySz \wedge z \models B))$. Then, along with the fact that $\mathcal{Y} \models xRy \wedge y \models A$, we can conclude that there exists a $z \in \mathcal{Y}$ such that $\mathcal{Y} \models xRz \wedge ySz \wedge z \models B$. As $L = x$, $\mathcal{Y} \models L = y \wedge xRz \wedge ySz$, by (S), we have that for all k, $\mathcal{Y} \models \mathrm{Con}(PA_k + L = z)$. So there is an end-extension $\mathcal{Z}$ of $\mathcal{Y}$ such that $\mathcal{Z} \models \mathrm{PA} + L = z$, by Theorem 13 above.

We will now show that $\mathcal{Z}$ models $\mathrm{PA} + B^*$. As '$z \models B'$ is a Σ_1-assertion that holds in $\mathcal{Y}$ and $\mathcal{Z}$ is an end-extension of $\mathcal{Y}$, $\mathcal{Z} \models (z \models B)$. Now, using the induction hypothesis: $\mathrm{PA} \vdash$ 'I is faithful on B' and reasoning as above, we have that $\mathcal{Z} \models$ 'I is faithful on B'. Combining this with $\mathcal{Z} \models L = z$ (construction of $\mathcal{Z}$) and $\mathcal{Z} \models (z \models B)$ (above) gives that $\mathcal{Z} \models B^*$. We assumed that no model of $\mathrm{PA} + A^*$ has an end extension which is a model of $\mathrm{PA} + B^*$, but we have found that $\mathcal{Y}$ does. Hence, $\mathrm{PA} + A$ really *does* interpret $\mathrm{PA} + B$. In other words C^*, and we are done.

Proposition 34. (ACA_0) *If* $x \not\models A \rhd B$ *and* $L = x$*, then* $\neg C^*$*.*

Continue working in ACA_0. Suppose that $x \in W$, $x \models \neg(A \rhd B)$ and $L = x$. We must now prove that $\neg(A \rhd B)^*$. Remember this denotes that $\mathrm{PA} + A^*$ does not interpret $\mathrm{PA} + B^*$ by how we defined the realization $*$. By Theorem 12, to prove $\neg(A \rhd B)$ it will be enough to find a model $\mathcal{Y}$ of $\mathrm{PA} + A^*$ that has *no* end-extension $\mathcal{Z}$ that models $\mathrm{PA} + B^*$. First note that the first and second suppositions jointly guarantee that there is a witnessing y in W such that

$$xRy \wedge y \models B \wedge \forall z \in W[(xRz \wedge ySz) \to z \models \neg B]. \tag{4.8}$$

We will now show that any model $\mathcal{Y}$ of the sentence $L = y$ will be a model of $\mathrm{PA} + A^*$ that has no end-extensions $\mathcal{Z}$ that model $\mathrm{PA} + B^*$. That such a model exists follows from $L = x$ (assumption), xRy (above), and property (R) jointly implying that $\mathrm{Con}(\mathrm{PA} + L = y)$. Hence $\mathrm{PA} + L = y$ has a model, say $\mathcal{Y}$ (by the formalized version of the soundness of PA). The induction hypothesis here is: $\mathrm{PA} \vdash$ 'I is faithful on A'. Reasoning as above concerning the induction hypothesis, we have that $\mathcal{Y} \models$ 'I is faithful on A'. Notice that as '$y \models A$' is a true Σ_1-assertion (i.e. it holds in the standard model of arithmetic), then it will be true in all models of PA, including $\mathcal{Y}$. So $\mathcal{Y} \models (y \models A)$. Therefore,

$$\mathcal{Y} \models L = y \wedge y \models A \wedge \text{`}I \text{ is faithful on } A\text{'}. \tag{4.9}$$

So, we have that $\mathcal{Y} \models A^*$ by faithfulness, and hence that $\mathcal{Y} \models \mathrm{PA} + A^*$.

We want to prove now that no end-extension of $\mathcal{Y}$ models PA + B^*. Suppose for contradiction that there *is* an end-extension $\mathcal{Z}$ of $\mathcal{Y}$ such that $\mathcal{Z} \models \text{PA} + B^*$. We first prove:

Claim 2. Suppose there is a $z \in \mathcal{Z}$ such that $\mathcal{Z} \models L = z$. Then $\mathcal{Z} \models z \in W \wedge xRz \wedge ySz$.

Proof: By construction of $\mathcal{Y}$, $\mathcal{Y} \models L = y$. By the specification of L as the limit of the function F, we have that $\mathcal{Y} \models L \in \text{Range}(F)$. So $\mathcal{Y} \models (y \in \text{Range}(F))$ by substitution from the previous equation. Since $\mathcal{Z}$ was assumed to be an end-extension of $\mathcal{Y}$, by Theorem 10, $\mathcal{Z} \models y \in \text{Range}(F)$. By the same theorem we have that $\mathcal{Z} \models xRy$.

Then from the assumption of the claim that $\mathcal{Z} \models L = z$ along with property (¬S) gives us that $\mathcal{Z} \models ySz$. Combining with above, we have $\mathcal{Z} \models xRySz$. By the property (¬ R), we have that $\mathcal{Z} \models xRz$. We also need to show that $z \neq 0$, and hence $z \in W \cup \{0\}$. We know that R is always increasing in the sense that if aRb, then $a < b$. Then as $x \in W \cup \{0\}$, $z \neq 0$. Hence $z \in W$ (not simply $z \in W \cup \{0\}$). This proves the claim that if there is a $z \in \mathcal{Z}$ such that $\mathcal{Z} \models L = z$, then:

$$\mathcal{Z} \models z \in W \wedge xRz \wedge ySz \tag{4.10}$$

Now to finish the proof, by our choice of y as the witness that $\mathcal{Y} \not\models A \rhd B$, we have that:

$$\forall w \in W[(xRw \wedge ySw) \rightarrow w \models \neg B] \tag{4.11}$$

This assertion is primitive recursive, so it much be satisfied in the model $\mathcal{Z}$. As the function F is formalizable in the model $\mathcal{Z}$ and F always has a limit, we have that for some $z \in \mathcal{Z}$, $\mathcal{Z} \models L = z$. We will force a contradiction by considering what happens at the node z. Apply equation (4.10) above to whatever this z is to obtain $\mathcal{Z} \models z \in W \wedge xRz \wedge ySz$. Applying formula (4.11) to z, we find that $\mathcal{Z} \models (z \models \neg B)$. Combine this with above to get:

$$\mathcal{Z} \models (L = z \wedge z \models \neg B) \tag{4.12}$$

Now consider the induction hypothesis: 'PA is faithful on B'. Again reason that this must hold in PA even though we first assume it outside of PA. Take the second faithfulness condition: 'if $z \models \neg B$ and $L = z$, then $\neg B^*$'. As this holds in PA, it must be satisfied by any model of PA, including $\mathcal{Z}$. Reason in the model $\mathcal{Z}$ using formula (4.12) to conclude that $\mathcal{Z} \models \neg B^*$.

Contradiction. We have assumed that $\mathcal{Z}$ is an end-extension of $\mathcal{Y}$ such that $\mathcal{Z} \models \text{PA} + B^*$. So $\text{PA} + A^*$ does not interpret $\text{PA} + B^*$. In other words, $\neg(A \rhd B)^*$, as desired. This completes Part 2, and we are done with the proof of Lemma 32 that $*$ if faithful on all modal formulas. □

Now we are in position to finish the proof of the Arithmetical Completeness Theorem:

(Outside of PA) If ILM $\nvdash A$, then by the Simplified Model Completeness Theorem 23, there is a ILM-model $\mathbf{W}$, with root b, such that $b \models \neg A$. As the model is provably primitive recursive, we also have that $\text{PA} \vdash b \models \neg A$. From the second faithfulness condition, we have:

$$\text{PA} \vdash \text{`}\forall x \in \mathbf{W}(x \models \neg A \wedge L = x) \to \neg A^{*}\text{'} \tag{4.13}$$

Working with the assumption above that $\text{PA} \vdash b \models \neg A$, we get $\text{PA} \vdash L = b \to \neg A^*$. Then by the deduction theorem we then have that

$$\text{PA} + L = b \vdash \neg A^* \tag{4.14}$$

By remark (31), for all x, $\text{PA}+L = x$ is consistent, and hence has a model. In particular, $\text{PA} + L = b$ has a model, say $\mathcal{M}$, such that $\mathcal{M} \models \text{PA} + L = b$. Then, dropping $L = x$ from formula (4.14), we have:

$$\mathcal{M} \models \neg\ A^* \tag{4.15}$$

But anything that a model of PA satisfies (e.g. A^*) cannot be disproved by PA. Hence $\text{PA} \nvdash A^*$, and we are done. We now have both the Arithmetical Soundness of ILM and the Arithmetical Completeness of ILM. Therefore we have the Main Result, and we are done. □

Acknowledgments: I would like to thank Professor Warren Goldfarb for serving as my advisor for this project. His role was indispensable, especially as an audience for presentation of this material. Who knew we would share almost a year together on provability and interpretability! I would also like to thank Professor Jeremy Avigad for materials and assistance in the very beginning of the project, and Professor Gerald Sacks for helpful discussion at the very end of it. I am also grateful to the Harvard College Research Program, which provided summer research funding to explore topics for this project. I thank my friends, who heard me talk about this project for the first half of the year, only to have me disappear in the second half while writing it. I especially thank my family for their advice and support. Their help in the last weeks of the project was invaluable.

Bibliography

[1] Berarducci, Alessandro, *The Interpretability Logic of Peano Arithmetic*, Journal of Symbolic Logic. **3** (1990), 1059–1089.

[2] Boolos, George, *The Logic of Provability*, Cambridge University Press, New York, 1993.

[3] de Jongh, Dick, and Veltman, Frank, *Provability Logics for Relative Interpretability*, Mathematical Logic, Plenum Press, New York, 1990, 1-19.

[4] Feferman, Soloman, *Arithmetization of metamathematics in a generalized setting*, Fundamenta Mathematicae. **49** (1960).

[5] Hinman, Peter, *Fundamentals of Mathematical Logic*, A K Peters Press, Wellesley MA, 2005.

[6] Japaridze, Giorgi, et. al., *The Logic of Provability*, Handbook of Proof Theory, Elsevier Science B.V., Amsterdam, 1998, 475-546.

[7] Kaye, Richard, *Models of Peano Arithmetic*, Oxford University Press, New York, 1991.

[8] Lindström, Per, *Aspects of Incompleteness, 2nd edition*, Association of Symbolic Logic, Natick, Massachusetts, 2003.

[9] Maker, David, *Model Theory: An Introduction*, Springer, New York, New York, 2002.

[10] Tarski, Alfred, et. al., *Undecidable Theories*, North-Holland, Amsterdam, 1953.

[11] Shavrukov, Volodya, *Logic of relative interpretability over Peano arithmetic*, Preprint No. 5, Steklov Mathematical Institute, Moscow, 1988.

[12] Visser, Albert, *Preliminary notes on interpretability logic*, Logic group preprint series, no. 29, University of Utrecht, Utrecht, 1988.

[13] Visser, Albert, *Interpretability Logic*, Mathematical Logic, Plenum Press, New York, 1990, pp. 175-209.

[14] Visser, Albert, *An overview of Interpretability Logic*, Advances in Modal Logic, Vol. 1 Berlin, **1** (1996), pp. 307-59.