An overview of interpretability logic

Visser, A.

Publication date
1997

Document Version
Submitted manuscript

Citation for published version (APA):
Visser, A. (1997). *An overview of interpretability logic*. (Research report Logic Group Preprint Series; No. 174). Department of Philosophy, University of Utrecht.

An overview of Interpretability Logic

Albert Visser[*]
Department of Philosophy, University of Utrecht
albert.visser@phil.ruu.nl

March 12, 1997

1 Introduction

A miracle happens. In one hand we have a class of marvelously complex theories in predicate logic, theories with 'sufficient coding potential', like PA (Peano Arithmetic) or ZF (Zermelo-Fraenkel Set Theory). In the other we have certain modal propositional theories of striking simplicity. We translate the modal *operators* of the modal theories to certain specific, fixed, defined *predicates* of the predicate logical theories. These special predicates generally contain an astronomical number of symbols. We interpret the propositional variables by arbitrary predicate logical sentences. And see: the modal theories are sound and complete for this interpretation. They codify precisely the schematic principles in their scope. Miracles do happen ....

Our miracle —as any good miracle— involves transsubstantiation. We translate between languages of incomparable signature. The modal languages do not contain quantifiers, the predicate logical languages do not contain modal operators. The modal operators can be translated to predicates because we transsubstantiate formulas occurring in the scope of a modal operator to closed terms (numerals) representing codes (gödelnumbers) of formulas of the target theory.

The miracle does not always work —as is to be expected of true miracles— we get no analogous result if we try to work with modal predicate logical languages. See [7].

Provability Logic studies formal provability as a modality employing ideas and methods of modal logic. Interpretability Logic extends Provability Logic by adding a binary modality. This modality can be given several interpretations. *Relative interpretability* and Π_1*-conservativity* are the most salient. (We will

[*] The author thanks Lev Beklemishev, Dick de Jongh, Carlos Areces and Stefan Iwan for invaluable comments and suggestions. Maarten de Rijke was the most supportive editor ever. I am grateful to Volodya Shavrukov for his graceful permission to include one of his unpublished results.

see some others.) Thus, Provability Logic and Interpretability Logic are part of a branch of Modal Logic where we do not study time, as in Temporal Logic, or obligation, as in Deontic Logic, but *formal theories*. An important philosophical difference is this. Time and obligation are not themselves mathematical objects. We model certain salient and interesting aspects of the central notions using classes of structures and study the interplay of the logics and the structures. Formal theories, in contrast, are themselves mathematical objects. They do not appear in the role of analysanda. The Kripke structures we employ play the role of technical auxiliaries, not analysantia.

This paper aims to survey the main results of Interpretability Logic. It does not pretend to be exhaustive. Below I give some reasons for studying Interpretability Logic. On a first reading this list of motivations could be very well skipped.

1.1 Beauty

Of course, there is the matter of beauty. However, beauty should not be advertized. Thus I will further pass it over in silence ....

1.2 Reasoning in the logics

Some non-trivial reasoning concerning interpretability can be formalized in the modal logics. The gain here is perspicuity and generality.[1]

Representing some substantial reasoning in modal terms was the main aim of Vı́těslav Švejdar in his classical paper [45]. Švejdar employs a combination of interpretability and witness comparisons for certain special formulations of proof predicates. The true potential of the ideas of Švejdar's paper is still not fully explored. It would certainly be benificial, from the didactical point of view, if some proofs concerning degrees of interpretability and the complexity of interpretability would be rephrased in Švejdar's language. For some of Švejdar's interpretations no arithmetical completeness results exist .

Another example of non-trivial reasoning in the logics is the alternative proof of a result of Solovay in [55]. See also section 6 of this paper.

1.3 Other spin-off

Let me first mention an example of spin-off of research into the question whether there is a ZF-sentence that is interpretable in ZF, but not in GB. Solovay's work on this question produced the method of shortening cuts. This method was e.g. used by Paris & Wilkie, by Pudlák and various others to prove several metatheorems. See [35],[37], [38]. The method was adapted by Nelson to build stronger and stronger theories in his predicativist programme. See [34].

[1] For the case of provability logic the programme of using the language to represent non-trivial reasoning was strongly advocated by Craig Smorynński. See his book [42].

The *expertise* developed in proving arithmetical completeness theorems for Interpretability Logic was used with good result by Shavrukov in the study of the combined logic for provability and a Feferman predicate. See [40].

Dick de Jongh and Duccio Pianigiani, in their[11], used the work of Hájek and Montagna ([23],[24]) to solve an open problem posed by Guaspari in [22].

1.4 Internal interest

The study of Interpretability has also some internal interest in Provability Logic considered as a project. It is this. Solovay's result turned out to be completely general. Consider any theory into which $I\Delta_0 + \mathsf{EXP}$ is interpretable, say by the interpretation $\mathcal{N}$. Suppose that we arithmetize 'in $\mathcal{N}$' and suppose that our theory is $\Sigma^0_1(\mathcal{N})$-sound. Then its provability logic is precisely Löb's Logic. In case we drop the condition of $\Sigma^0_1(\mathcal{N})$-soundness, we get only relatively uninteresting well understood variants of Löb's Logic. See [51]. There are only two ways to escape the stability (keeping modal language and the interpretation of the box as provability fixed). The first is to go below $I\Delta_0 + \mathsf{EXP}$, to weaker theories, where Löb's Logic is still arithmetically valid, but where *we do not know whether it is complete*. See e.g. [6]. The second is to vary the logic. It is well known that if we consider the provability logic of Heyting Arithmetic[2], HA, we find a new, rich, weird and wonderful landscape of wild and surprising modal principles. See e.g. [59].

Of course, extending the modal language might also be a way to escape the stability of Solovay's result. It turns out that for Interpretability we do get good modal logics and Solovay style completeness results, but that we do not have the absolute stability of Provability Logic. There are two major classes of theories that have quite different interpretability logics. The first class is that of the sequential, Σ^0_1-sound, finitely axiomatized theories containing $I\Delta_0 + \mathsf{SUPEXP}$. Examples are: $I\Delta_0 + \mathsf{SUPEXP}$, $I\Sigma_n$ $(n > 0)$, ACA_0, GB. Theories in this class are sound and complete for the logic ILP. See [55]. The second class is that of sequential, locally essentially reflexive theories containing $I\Sigma_1$. Examples are PA and ZF. Theories in this class satisfy are sound and complete for the logic ILM. This result was proved independently by Alessandro Berarducci and Volodya Shavrukov. See [5] and [39]. Outside of these major classes we know very little. See section 9 and appendix B.

1.5 Philosophical interest

The philosophical interest of Provability Logic is that it analyzes Gödelian metamathematical reasoning in its bare essence. I think that this, all by itself, constitutes a substantial contribution. However, there is a bonus. The contrast between *provability* in some appropriate intuitive sense and *formal provability*

[2] Heyting Arithmetic is, in esence, Peano Arithmetic with intuitionistic logic instead of classical logic.

is shown most strikingly in the comparison of the modal systems $\mathsf{S4}$ and Löb's Logic, GL, corresponding to these notions. Specifically, comparison between Reflection and Löb's Principle seems a potent antidote to the misguided impression that the difference between the notions is one of strength, the impression that Gödel's Theorem means that Human Mental Powers exceed what formal systems can do, the Myth of the Mental Muscles. The difference between the two notions is, I submitt, rather one of *kind*. Their comparison in terms of strength is as absurd as comparing the strength of a master of Chess and one of Karate.

Interpretability Logic does not add —as far as I can see— anything along the lines of the above sources of interest over and above what we already had in Provability Logic. It *does* add the expressive power to reflect modally the consequences, not only of Gödelian Incompleteness, but also of Gödelian Completeness in the form of the interpretation version of the Model Existence Lemma. It seems to me that the results of my paper [57] throw some light on Nelson's program for founding predicative mathematics using interpretations, see [34]. Specifically, the result that $I\Delta_0 + \mathsf{EXP}$ is equivalent modulo interpretability with $I\Delta_0 + \Omega_1 + \mathsf{Con}(I\Delta_0 + \Omega_1)$, seems to show that the insight in the consistency of his own theories must be suspect for the predicativist who rejects the cogency of the totality of exponentiation. (See also subsection 11.2.)

1.6 Selection of notions (?)

Interpretability Logic studies abstract global properties of certain arithmetical predicates of independent interest. We could reverse the direction and use principles of Interpretability Logic as a check list to see whether a candidate notion, e.g. for comparing theories, is a reasonable one. I have two tentative examples. The first is the notion of Σ-preservativity, which is proposed as a metamathematical tool in the study of Heyting's Arithmetic, HA, and its extensions. See [52], [59] and subsection 10.4 of this paper. The second is the formula of the Friedman Characterization of interpretability in sequential, finitely axiomatized theories applied to infinitely axiomatized theories. This would give us a way of comparing theories that is sensitive to the ease in which a theory can prove concrete Δ^0_1-sentences.[3]

2 A list of theories and notions

At this point we interpolate a little list of notions and theories. The best reading strategy is to glance through it and to return to it when needed. A good book where most of the notions and theories mentioned here are treated is [25].

[3] I have a sketch of a proof that $I\Delta_0 + \{\mathsf{supexp}(\underline{n})\!\downarrow \mid n \in \omega\}$ is equivalent w.r.t. this notion with $I\Delta_0 + \mathsf{EXP}$. This in spite of the fact that our first theory clearly proves the same theorems as $I\Delta_0$.

- The language of arithmetic —unless stated otherwise— is the language of 0, S (successor), $+$ and $\times$. If Γ is a set of arithmetical formulas, closed under subformulas and substitution of terms, then $I\Gamma$ is the theory containing the basic facts about $0, S, +, \times$, plus induction restricted to Γ. Two salient theories are $I\Delta_0$ and $I\Sigma_1$. PA is Peano Arithmetic, the theory of full induction.

- We write $\#A$, $\#t$ for the *gödelnumbers* of A, respectively t. For any number n let $\underline{n}$ be the *numeral* of n. In the context of weak theories one almost always employs numerals that correspond to binary notations. Let num be the function mapping n to $\#\underline{n}$. Par abus de langage we also use 'num' for the arithmetization of num. We write $*$ both for concatenation and its arithmetization

- Prov_T is the arithmetization of provability in T. We write $\Box_T A$ for $\mathsf{Prov}_T(\underline{\#A})$. We also use the notation $\Box_T A$ in case A containes free variables. What we mean here is best explained by example. $\Box_T x = y$ means $\mathsf{Prov}_T(t(x,y))$, where $t(x,y)$ is the term $\mathsf{num}(x) * \underline{\#=} * \mathsf{num}(y)$.

- We will write $\Diamond$ for $\neg\Box\neg$. So e.g. $\Diamond_T\top$ will be another way of writing $\mathsf{Con}(T)$.

- Ω_1 is the axiom expressing that the function $x^{\mathsf{entier}((^2\mathsf{log}(x))^2)}$ is total. Here, $\mathsf{entier}(x)$ is the largerst natural number $\leq x$. EXP expresses that exponentiation is total, SUPEXP that superexponentiation is total, etc. One can show that these axioms can be formulated using Π_2-formulas of the usual arithmetical language.

- $I\Delta_0 + \Omega_1$ is studied by Paris and Wilkie in their [35]. It is a natural theory to do arithmetization in up to the formalization of Gödel's Second Incompleteness Theorem. Complicated Rosser arguments —like the proof of Solovay's Theorem— may present difficulties.

- S^1_2 is a theory introduced by Buss in his [8]. It is weaker than $I\Delta_0+\Omega_1$ and can be consider as the theory of $\mathsf{P\text{-}TIME}$. Its provably recursive functions are precisely the $\mathsf{P\text{-}TIME}$ computable ones. In S^1_2 we can formalize the Second Incompleteness Theorem.

- $I\Delta_0 + \mathsf{EXP}$ is also known as Elementary Arithmetic or EA. Its provably recursive functions are precisely the Kalmar elementary functions. All Rosser style arguments can be formalized in this theory.

- In $I\Delta_0 + \mathsf{SUPEXP}$ one can formalize cut-elimination for Predicate Logic.

- PRA is usually formulated as a theory with symbols for all primitive recursive functions and as axioms the appropriate defining equations corresponding to these symbols, plus induction for atomic formulas. Since this formulation of PRA does not quite fit our framework —the set of atomic symbols of its language being infinite— we will often think of different, but equivalent formulations. Two such formulations are:

- $I\Delta_0 + \mathsf{EXP} + \mathsf{SUPEXP} + \mathsf{SUPSUPEXP} + \ldots$.
- $I\Delta_0 + \mathsf{EXP}$ plus the Σ^0_1-induction rule. This rule states that if we have proved the premiss of Σ^0_1-induction, then we may draw its conclusion. We will call this theory $R\Sigma^0_1$.[4]

For an extensive discussion of formulations of PRA and related theories, see [2].

- Rfn_T is the *local reflection principle* for T, i.e. the schema $\Box_T A \to A$, for the sentences A of $\mathcal{L}_T$. RFN_T is the *global* or *uniform reflection principle* for T, i.e. the schema $\forall \vec{x}\, (\Box_T A(\vec{x}) \to A(\vec{x}))$, for formulas $A(\vec{x})$ of $\mathcal{L}_T$.
- ZF is Zermelo-Fraenkel Set Theory. GB is Gödel-Bernays Set Theory. GB is finitely axiomatized. ACA_0 is a finitely axiomatized extension of PA with classes. In many respects ACA_0 is to PA as GB is to ZF.

3 What is relative interpretability?

There is no analogue of the Church-Turing Thesis for interpretability. For one thing interpretations are dependent on the notion of Formal Provability, which is itself an artifact of the mathematical imagination. For another, the boundaries of what we count as an interpretation seem to be quite interest dependent.

It is clear that an interpretation of a theory V in a theory U should at least deliver a function f, from the sentences of V to the sentences of U, such that, for all sentences A of V, $V \vdash A \Rightarrow U \vdash \mathsf{f}(A)$ $(*)$.[5] Let, for example, U be PA and let V be ZFC. PA contains sufficient coding machinery to represent $\mathsf{Prov}_{\mathsf{ZFC}}(x)$. The mapping $A \mapsto \mathsf{Prov}_{\mathsf{ZFC}}(\underline{\#A})$ will satisfy $*$. Nobody however would count it as an interpretation in any sense. One objection against allowing this as an interpretation would be that we cannot use this mapping to establish relative consistency results. This objection, if accepted, would also rule out e.g. the Friedman Translation as an interpretation.[6]

Another proposal is to demand that interpretations commute with certain connectives. This proposal would rule out e.g. forcing, realizability, the double

[4] One can show that the Σ^0_1-induction rule is equivalent over $I\Delta_0 + \mathsf{EXP}$ to the Π^0_2-induction rule. Warning: this result does not hold over arbitrary extensions of $I\Delta_0 + \mathsf{EXP}$.

[5] Ironically, our definition of relative interpretation will not even fit this precise format. However, we can tell a story to explain that we only deviate in an unessential sense, that has to do with the details of the treatment of variables and thinking modulo α-conversion.

[6] To demand that interpretations commute with $\perp$ does not guarantee the possibility of relative consistency proofs. We could need the insight of relative consistency to show that $*$ holds. We get an example of this phenomenon by replacing $\mathsf{Prov}_{\mathsf{ZFC}}$ in our example by a predicate representing Feferman provability in ZFC.

negation translation. (The Friedman Translation only fails to commute with $\bot$.)

Rather than pursuing the problem of finding a reasonably general notion of interpretation satisfying certain intuitive constraints, we will study one given notion of interpretation: *relative interpretability.* This notion is due to Tarski, Mostowski and Robinson. See [47]. Roughly we demand that our interpretations commute with all the propositional connectives (including the all important $\bot$) and with the quantifiers modulo relativization to a domain. Moreover, we restrict ourselves to theories in classical Predicate Logic. This choice means a restriction on the generality of the results discussed in this paper. However, sometimes the results are stable under extension. E.g., in studying finitely axiomatized extensions of ZF in the language of ZF we could easily extend our notion to include forcing, without changing the corresponding Interpretability Logic.

From this point on we will confuse 'interpretation' with 'relative interpretation' in the precise sense defined below.

3.1 Defining relative interpretation

An important thing to remember is that we are interested in formalization of facts concerning interpretability in a wide range of theories. This means that our definition of interpretation must be relatively simple and managable. Moreover, all kinds of details that we usually abstract away from, like the precise choice of auxiliary variables, may become relevant. (Indices of such variables have to be coded too, so unhappy choices may produce codes that are to large for a weak theory to handle.) The definitions of the translation based on an interpretation vary across papers. The definition given here is, I think, the most convenient one. It has the advantage of handling function symbols of the interpreted language with relative ease.[7] It sidesteps, by the use of 'fresh' auxiliary variables and by avoiding substitution the hairy issue of variable clashes. Since unfortunate choices of auxiliary variables can cause undesirable and ugly growth of the translations, we will firmly regiment the use of the auxiliary variables by assigning each argument place its unique auxiliary representative, thus keeping our algorithm within linear time. Consider first order theories U and V with languages respectively $\mathcal{L}_U$ and $\mathcal{L}_V$. We assume that identity occurs in these languages and that we have only finitely many relation and function symbols. Constants are treated as 0-ary function symbols. We extend $\mathcal{L}_U$ with new variables $a_0, \cdots, a_n$, where n is the maximum of the arities of the relation and function symbols of $\mathcal{L}_V$. The a_i will be used to handle the machinery of

[7] Usually one defines interpretations only for relational languages, justifying this restriction by the existence of an algorithm to eliminate function symbols. Here we incorporate such an algorithm in the very definition of the translation.

the argument places in the translation. Say the resulting language is $\mathcal{L}_U^+$.[8] An interpretation $\mathcal{M}$ of V into U is given by two things:

- a function F mapping the relation symbols R and the function symbols f of $\mathcal{L}_V$ on formulas of $\mathcal{L}_U^+$. We demand:
  - if the arity of R is k, then the free variables of $F(R)$ are among $a_1, \cdots, a_k$,
  - if the arity of f is ℓ, then the free variables of $F(f)$ are among $a_0, \cdots, a_\ell$.

  (We could allow extra free variables ('parameters') in $F(R)$ and in $F(f)$. This would only cause some minor changes in the set-up.)
- a formula δ, with only a_0 free, of $\mathcal{L}_U^+$ giving the domain of the interpretation.

$\mathcal{M}$ gives a translation $(.)^{\mathcal{M}}$ of $\mathcal{L}_V$ in $\mathcal{L}_U^+$ in the following way. We write $A[v]$, for v a variable of $\mathcal{L}_U^+$ distinct from a_0, as an abbreviation of $\exists a_0(A \wedge a_0 = v)$. The translation of a formula A will have the same free variables as A itself. The translation of a term t will have as free variables the free variables of t plus the auxiliary variable a_0, which stands here for *the value of* t.

- $R(t_1, \cdots, t_k)^{\mathcal{M}} := \exists a_1 \cdots \exists a_k\, (F(R) \wedge (t_1)^{\mathcal{M}}[a_1] \wedge \cdots \wedge (t_k)^{\mathcal{M}}[a_k])$,
- $f(t_1, \cdots, t_\ell)^{\mathcal{M}} := \exists a_1 \cdots \exists a_\ell\, (F(f) \wedge (t_1)^{\mathcal{M}}[a_1] \wedge \cdots \wedge (t_\ell)^{\mathcal{M}}[a_\ell])$ (note that a constant c simply goes to $F(c)$, a formula containing only a_0 free),
- $x^{\mathcal{M}} := (a_0 = x)$,
- $(.)^{\mathcal{M}}$ commutes with the propositional constants,
- $(\forall x\, A)^{\mathcal{M}} := \forall x(\delta[x] \to A^{\mathcal{M}})$, $(\exists x\, A)^{\mathcal{M}} := \exists x(\delta[x] \wedge A^{\mathcal{M}})$.

Note that the usual algorithm for eliminating function symbols, is a special case of an interpretation in our sense. Simply put: $F(R) := R(a_1, \cdots, a_k)$ and $F(f) := f(a_1, \cdots, a_\ell) = a_0$. We will call this interpretation ID, since it fulfills the role of the indentity interpretation.

One can avoid the device of using '$A[t]$' by a slightly more elaborate use of a larger set of auxiliaries. I leave such variants to the fantasy of the reader. A more radical alternative is to develop syntax on sharing graphs and code the (labeled) graphs arithmetically. A more detailed discussion is outside the scope of this paper.

[8] The critical reader may object here, that by extending $\mathcal{L}_U$ to $\mathcal{L}_U^+$, we, in effect, extend U to U^+, thus changing the interpreting theory. This line of thought can lead to a fascinating discussion on theory individuation. Rather than entering into this discussion, let me point out that the auxiliary variables are easily eliminated by α-conversion from the translation. The gain of having the auxiliaries is just conceptual perspicuity.

Let $\delta_A := \bigwedge\{\delta[x] \mid x \text{ free in } A\}$. U interprets V via $\mathcal{M}$ if: for all theorems (not necessarily sentences) A of V, $U \vdash \delta_A \to A^{\mathcal{M}}$. Alternatively, we can use axioms instead of theorems in the definition. However, we must note two things here. First, we must include statements expressing the functionality of the function symbols among the axioms. Secondly in weak theories like $I\Delta_0$, or even PRA, we do not have Σ_1-collection[9]. This lack blocks the derivation of the equivalence of 'theorems-interpretability' and 'axioms-interpretability'. The delicate points concerning formalization in a weak environment are treated extensively in [56]. We will ignore these subtleties here. We write:

- $\mathcal{M} : U \rhd V \Leftrightarrow U$ interprets V via $\mathcal{M}$,
- $U \rhd V :\Leftrightarrow$ for some $\mathcal{M}$, $\mathcal{M} : U \rhd V$,
- $U \equiv V :\Leftrightarrow U \rhd V$ and $V \rhd U$,
- $A \rhd_U B :\Leftrightarrow (U + A) \rhd (U + B)$. (We say: A interprets B over U.)

We can view theories and interpretations as objects and morphisms of a category. We did not build in into the notion of interpretation any data concerning theories. So strictly speaking a morphism in the category of theories and interpretations is a triple $\langle V, \mathcal{M}, U\rangle$, such that $\mathcal{M} : U \rhd V$.[10]

A closely related notion is local interpretability. Let $\mathsf{FS}(V)$ be the set of subtheories of V which are axiomatized by finitely many axioms of V. We define local interpretability as follows.

- $U \rhd_{\mathsf{loc}} V :\Leftrightarrow \forall V_0 \in \mathsf{FS}(V)\, \exists \mathcal{M}_0,\ \mathcal{M}_0 : U \rhd V_0$,
- $U \equiv_{\mathsf{loc}} V :\Leftrightarrow U \rhd_{\mathsf{loc}} V$ and $V \rhd_{\mathsf{loc}} U$,
- $A \rhd_{\mathsf{loc},U} B :\Leftrightarrow (U + A) \rhd_{\mathsf{loc}} (U + B)$. ($A$ locally interprets B over U.)

If want to emphasize the contrast with local interpretability, we will call ordinary interpretability *global interpretability*. It is well known that local and global interpretability do not coincide. E.g. let $\mathsf{Con}_n(\mathsf{GB})$ be the consistency statement for GB w.r.t. proofs in which no formulas occur of complexity greater than n. Then:

$$\mathsf{GB} \rhd_{\mathsf{loc}} (I\Delta_0 + \{\mathsf{Con}_n(\mathsf{GB}) \mid n \in \omega\}),$$

but *not*: $\mathsf{GB} \rhd (I\Delta_0 + \{\mathsf{Con}_n(\mathsf{GB}) \mid n \in \omega\})$. An advantage of local interpretability is that it is less complex (Π^0_2) than global interpretability (Σ^0_3). Some basic

[9] Σ_1-collection is the principle $\forall x \exists y\, A \to \forall a \exists b \forall x {\leq} a \exists y {\leq} b\, A$, where A is Σ_1 and a, b do not occur in A.

[10] We take the morphisms to go from interpreted theory to the interpreting theory, since this convention is in consonance with the tradition in the study of the *degrees* of interpretability to have the *stronger* theories in the *higher* degrees. By erasing 'morphism identity' we get a preorder; by dividing out the induced equivalence relation we get the usual partial order of degrees.

facts on the relationship between local and global interpretability are presented in appendix C. For a broader discussion on local interpretability and a description of some further notions of reduction, the reader is referred to [33].

3.2 Interpretations as internal models

The construction of a model of two dimensional elliptic geometry on the sphere can be considered as the construction of a model of two dimensional elliptic geometry inside a model of three dimensional Euclidean geometry. We will say that the first model is an *internal* model of the second. Interpretations appear in the literature almost always as internal models. The reason for this preference is clear: an internal model can be visualized. An interpretation gives *a uniform method of assigning internal models.*[11] This mapping is described in some detail below. Relative consistency proofs employing interpretations do not need to talk about models at all. E.g. the statement that *if* ZF *is consistent, then* PA *is consistent* can be verified in weak theories like S^1_2 and $I\Delta_0 + \Omega_1$.

For completeness we describe the mapping of models associated with an interpretation. Consider a model $\mathcal{K} = \langle K, I\rangle$ of U. Suppose $\mathcal{M} : U \rhd V$. We write $[a : k]$ for 'the assignment that sends a to k'. (I do not want to be specific here on the question whether assignments have to defined on all variables or not. In the first case we need some convention of what happens with the non-displayed variables.) Define a new model $\mathcal{K}^{\mathcal{M}} := \mathcal{N} := \langle N, J\rangle$, as follows.

- $k \approx k' :\Leftrightarrow \mathcal{K}, [a_1 : k, a_2 : k'] \models F_{\mathcal{M}}(=)$,
- $N_0 := \{k{\in}K \mid \mathcal{K}, [a_0 : k] \models \delta_{\mathcal{M}}\}$,
- Let $k \in N_0$. Then $[k] := \{k'{\in}N_0 \mid k \approx k'\}$,
- $N := \{[k] \mid k \in N_0\}$,
- Let ℓ be the arity of R.
  $J(R) := \{\langle [k_1], \cdots, [k_\ell]\rangle \mid \mathcal{K}, [a_1 : k_1, \cdots, a_\ell : k_\ell] \models F_{\mathcal{M}}(R)\}$
- Let ℓ be the arity of f.
  $J(f) := \{\langle\langle [k_1], \cdots, [k_\ell]\rangle, [k_0]\rangle \mid \mathcal{K}, [a_0 : k_0, a_1 : k_1, \cdots, a_\ell : k_\ell] \models F_{\mathcal{M}}(f)\}$.

Clearly $\mathcal{N}$ will be a model of V. It is rather striking that the examples of 'models' that stood at the cradle of modeltheory, the alternative interpretations of geometry, can all be viewed as interpretations.[12] In the next subsection we provide some salient examples of interpretations.

[11] See [32] for a paper in which interpretations are studied as operators on models.

[12] Let $\mathsf{Mod}(U)$ be the set of models satisfying U. Then, $\mathcal{K} : U \rhd V$ yields a function from $\mathsf{Mod}(U)$ to $\mathsf{Mod}(V)$. Considering $\mathcal{K}$ as a morphism $V \xrightarrow{\mathcal{K}} U$, we see that Mod can be viewed as *a contravariant functor* from the category of theories and interpretations to the category of sets. It is easy to provide examples to show that even if $\mathcal{K} : U \rhd V$ is *faithful*, i.e. if the set of interpreted theorems coincides with the theorems of V, still the associated function between the models sets need not be surjective.

3.3 Examples of interpretations

Interpretations are everywhere dense!

1. The interpretation of two-dimensional elliptic geometry in Euclidean geometry of three dimensions. In this interpretation polar points on the sphere are identified. Thus the interpretation of identity is a non trivial equivalence relation. The interpretation employs free parameters, since we need to choose an abitrary centre and diameter for our sphere.

2. The Poincaré model and the Beltrami-Klein model of hyperbolic geometry. In these models the points of two dimensional hyperbolic geometry are interpreted as the points of the interior of a circle C. The lines are interpreted in the Poincaré model as diameters and segments of circles orthogonal to C. In the Beltrami-Klein model lines are simply segments of lines. The Poincaré model is faithful with respect to angles. E.g. the well known theorem that, in hyperbolic geometry, the sum of the angles of a triangle is strictly less than 180°, can be seen in one glance using Euclidean intuitions. See e.g. [27], p227. Alternatively, see [21]. In the Beltrami-Klein model all kinds of facts concerning incidence can be seen immediately. For example the fact that there alway is a line (asymptotically) parallel to both rays of an angle: in the model this will be the line connecting the points of intersection of the rays with the circle C. Thus the Euclidean models of hyperbolic geometry have in addition to *foundational importance*, also *heuristic value*.

3. The interpretation of arithmetic in set theory. This interpretation has *foundational importance*: it shows that numbers can be reduced to sets.

4. Gödel's interpretation of $\mathsf{ZF}+(\mathsf{V}=\mathsf{L})$ in ZF. This interpretation provides a relative consistency proof of $\mathsf{ZF}+\mathsf{CH}$ w.r.t. ZF.

5. The interpretation of elementary syntax in arithmetic. This interpretation plays a central role in the verification of Gödel's First Incompleteness Theorem and in both statement and proof of Gödel's Second Incompletenss Theorem.

6. The interpretation of $I\Delta_0 + \mathsf{Con}(\mathsf{ZF})$ in GB. This interpretation gives us a metamathematical lemma, from which we may conclude superexponential speed-up of GB over ZF. See [38].

7. The interpretation of $I\Delta_0$ in Robinson's arithmetic Q. This interpretation plays an important role in the development of Predicative Arithmetic in a foundational program worked out by E. Nelson. See [34].

8. Consider a sequential theory U, i.e. a theory containing a sufficient amount of machinery to handle sequences of arbitrary objects of the theory. Let

U contain a modicum of arithmetic (e.g. Robinson's arithmetic plus the axioms expressing that the standard ordering on the natural numbers is linear). Suppose $U \vdash \mathsf{Con}(V)$. Then $U \rhd V$. The interpretation is constructed by mimicking the Henkin model construction in U. Where we lack induction, we employ definable cuts using Solovay's method of shortening cuts to 'load' our cuts with some additional desirable properties. (See [56] or [57] for a careful exposition.)

9. A nonexample: Suppose PA is formalized using $\bot$, treating $\neg A$ as an abbreviation of $(A \to \bot)$. Let PA_0 be the theory obtained by replacing $\bot$ by $0 = 1$ in the non-induction axioms. Define a 'pseudo-interpretation' $\mathcal{M}$ as: *replace* $\bot$ *by* $0 = 1$. Then: $\mathcal{M} : \mathsf{PA}_0 \rhd \mathsf{PA} \supseteq (\mathsf{PA}_0 + \mathsf{Con}(\mathsf{PA}_0))$. So PA_0 validates: $\top \rhd \Diamond\top$, for the extended sense of interpretation where we allow $\bot$ to go to some sentence.

Dear reader, undoubtedly you miss your favourite example of an interpretation in the list. Please add it, 'in thought'.

3.4 Interpretations and arithmetic

We consider theories U formalized in predicate logic with reasonably simple axiom sets. A plausible demand is that these axiom sets are Σ^b_1. We ask that a suitable weak theory of arithmetic is interpretable in U. A good choice is $I\Delta_0 + \Omega_1$ or, alternatively, Buss' S^1_2. For information about weak theories, see e.g. [35],[8],[25]. The usual arithmetization of syntax, leading up to Gödel's Second Incompleteness Theorem, can be formalized in $I\Delta_0 + \Omega_1$ or S^1_2. We will always code syntax in the natural numbers of the theory.

Strictly speaking we are considering pairs $\langle U, \mathcal{N}\rangle$, where $\mathcal{N}$ is the designated interpretation of the natural numbers. The point is important, since one can always define different systems of natural numbers that are not provably isomorphic. x, y, z, u, v always range over these designated natural numbers. Here is an example of one theory with different designated sets of natural numbers.

Example 3.1 $I\Delta_0 + \mathsf{SUPEXP}$ verifies that GB is conservative over ZF w.r.t. the language of ZF. So a weak theory knows that GB and ZF are equiconsistent. Let ω be the usual interpretation of the natural numbers in GB. Clearly, $\langle \mathsf{GB}, \omega\rangle \vdash \mathsf{Con}(\mathsf{GB}) \leftrightarrow \mathsf{Con}(\mathsf{ZF})$, and, hence, by the Second Incompleteness Theorem, $\langle \mathsf{GB}, \omega\rangle \nvdash \mathsf{Con}(\mathsf{ZF})$. On the other hand one can find an interpretation $\mathcal{I}$, such that $\langle \mathsf{GB}, \mathcal{I}\rangle \vdash I\Delta_0 + \Omega_1 + \mathsf{Con}(\mathsf{ZF})$. Note that we cannot have, on pain of contradiction, $\langle \mathsf{GB}, \mathcal{I}\rangle \vdash \mathsf{SUPEXP}$. Closer inspection shows that we cannot even have $\langle \mathsf{GB}, \mathcal{I}\rangle \vdash \mathsf{EXP}$. ❑

The example shows that talk like GB *and* ZF *have the same strength* is somewhat misleading: it depends on how one compares. In practice we will leave the designated set of natural numbers implicit: they will always be clear from the context.

One of the nice things of interpretability logic is that it will enable us to make distinctions between certain *kinds* of theories. We introduce two important notions: *sequentiality* and *reflexivity.* A theory is *sequential* if we can define/code sequences of arbitrary objects from the domain of the theory in the theory. A place in a sequence or the length of a sequence is taken in the designated natural numbers of the theory. For a full discussion see e.g. [25].

- A theory is *reflexive* if it proves the consistency of each of its finite subtheories.
- A theory is *locally essentially reflexive* if all its finite sentential extensions are reflexive. Alternatively, a theory is *locally essentially reflexive* if it proves the local reflection principle for each of its finite subtheories. So if T is our theory and if we write $\Box_{T,n}$ for the arithmetization of provability from the first n axioms of T, then T is locally essentially reflexive if, for all sentences $A{\in}\mathcal{L}_T$ and for all n, $T \vdash \Box_{T,n}A \to A$.
- A theory is *globally essentially reflexive* or *uniformly essentially reflexive* or, simply, *essentially reflexive* if it proves the uniform reflexion principle for all its finite subtheories. So T is essentially reflexive if, for all formulas $A(\vec{x}){\in}\mathcal{L}_T$ and for all n, $T \vdash \forall\vec{x}(\Box_{T,n}A(\vec{x}) \to A(\vec{x}))$.
- A theory is *verifiably* reflexive , etcetera, if it verifies the formalized statement of its own reflexivity, etcetera. E.g. the formalized version of local uniform reflexivity for T is: $\forall A{\in}\mathsf{Sent}_{\mathcal{L}_T}\ \forall x\ \Box_T(\Box_{T,x}A \to A)$.

For an extensive discussion of reflection principles, see [1],[2],[3]. By the Second Incompleteness Theorem, locally reflexive theories cannot be finitely axiomatizable. If a theory is sequential and satisfies full induction, then it will be verifiably uniformly essentially reflexive. Conversely, if a uniformly essentially reflexive theory extends $I\Delta_0 + \Omega_1$, then it satisfies full induction. So for sequential theories extending $I\Delta_0 + \Omega_1$:

> Induction = uniform essential reflexivity = verifiable uniform essential reflexivity.

Consider any theory T. If we extend T with its own local reflection principle, then the resulting theory, say U, will be locally essentially reflexive. By a result of Feferman, U will be contained in T plus the true Π^0_1-sentences. $I\Sigma_1$ is not reflexive. PRA is verifiably reflexive but not essentially so. PA, ZF, ZFC are verifiably uniformly essentially reflexive.

Let Γ be a set of sentences present, possibly via a fixed interpretation, both in the language of T and U. We say that T is Γ-*conservative* over U iff for all A in Γ, $U \vdash A \Rightarrow T \vdash A$. We write:

- $T\ \rhd_{\Gamma\text{-}\mathsf{cons}} U :\Leftrightarrow T$ is Γ-*conservative* over U,

- $A \rhd_{\Gamma\textsf{-cons},T} B :\Leftrightarrow (T+A) \rhd_{\Gamma\textsf{-cons}} (T+B)$.

The use of conservativity to compare theories looms large in the literature.

For some kinds of theories we have pleasant characterizations of interpretability. We write $\mathsf{Con}_n(U)$ for the consistency of all axioms of T with gödelnumber smaller or equal than n.

Fact 3.2 Suppose T is reflexive. Then we have the following.

1. $T \rhd U \Leftrightarrow T \rhd_{\mathsf{loc}} U$
2. $T \rhd U \Leftrightarrow$ for all n $T \vdash \mathsf{Con}_n(U)$
3. Suppose that U is reflexive and satisfies sentential Σ^0_1-completeness. A sufficient condition for full Σ^0_1-completeness is the presence of the axiom EXP. We have: $T \rhd U \Leftrightarrow T \rhd_{\Pi_1\textsf{-con}} U$
4. Suppose that T is T-verifiably locally essentially reflexive and that it proves full Σ-completeness. Then:

$$\begin{array}{rcl} T \vdash A \rhd_T B & \leftrightarrow & A \rhd_{\mathsf{loc},T} B \\ & \leftrightarrow & A \rhd_{\Pi_1\textsf{-con},T} U \\ & \leftrightarrow & \forall x\ \Box_T(A \to \Diamond_{T,n} B) \end{array}$$

❑

The equivalence 3.2(2) is the important Orey-Hájek characterization. We elaborate on the proof in appendix C. Note that the notions of reflexivity, Π_1-conservativity and satisfying the Orey-Hájek characterization do depend on the designated numbers of our theories, but that interpretability and local interpretability do not. For extensions of PA in the arithmetical language we have a purely model theoretical characterization of interpretability (and ipso facto Π_1-conservativity).

Fact 3.3 Let T, U be extensions of PA in the language of PA. Then:

$$T \rhd U \Leftrightarrow \text{ all } \mathcal{M} \text{ with } \mathcal{M} \models T \text{ have end-extension } \mathcal{N} \text{ with } \mathcal{N} \models U.$$

❑

We end this section with the Friedman characterization of interpretability for finitely axiomatized sequential theories. See, for a proof, [55].

Fact 3.4 Let T, U be finitely axiomatized sequential theories. We write $\triangle$ for cut-free/tableaux/Herbrand provability. Let $\nabla := \neg\triangle\neg$. Remember that $\mathsf{EA} = I\Delta_0 + \mathsf{EXP}$. We have: $\mathsf{EA} \vdash T \rhd U \leftrightarrow \Delta_{\mathsf{EA}}(\nabla_T \top \to \nabla_U \top)$. ❑

It is an open question to give a characterization of interpretability that works for all sequential theories that contain a sufficient amount of arithmetic.

3.5 Excursion: oreysentences

Hilbert, in one of his more confused stages, suggested that truth is consistency. One problem with this suggestion is its lack of compliance with the law of non-contradiction: both A and $\neg A$ might be consistent. If we replace consistency in Hilbert's idea by interpretability, as Nelson seems to do, then the same problem emerges: for many theories U we can find a sentence O such that $\top \rhd_U O$ and $\top \rhd_U \neg O$. Such a sentence O is called an oreysentence. We provide some examples.

1. Let $\triangle$ stand for tableaux-provability (a version of cut-free provability). Then the gödelsentence of $\triangle_{I\Delta_0+\Omega_1}$ is an oreysentence. This was verified in Marianne Kalsbeeks masters thesis [29].

2. Let T be sequential. A rossersentence R of $\triangle_T$, tableaux provability in T, is a sentence such that $I\Delta_0 + \Omega_1 \vdash R \leftrightarrow \Delta_T\neg R \prec \Delta_T R$. Any rossersentence of $\triangle_T$ is an oreysentence of T. This is immediate by the Friedman characterization and the verifiability of Rosser's Theorem in EA.

3. Suppose T is reflexive. We define Feferman provability $\Box^*_T$ for T, as follows: $\Box^*_T A :\leftrightarrow \exists x(\Box_{T,x} A \wedge \Diamond_{T,x}\top)$. Here $\Box_{T,x}$ is provabilty from the first x axioms of T. One can show: $T \vdash \Diamond^*\top$ and $\Diamond^* A \rhd_T A$. The gödelsentence G of $\Box^*_T$ is an oreysentence of T, by the following reasoning. Argue in T. Suppose G. Then we have $\neg\Box^* G$, and, hence, $\Diamond^*\neg G$. So we can construct an interpretation $\mathcal{K}$ such that $(\neg G)^{\mathcal{K}}$. On the other hand if $\neg G$, then the identity interpretation ID will give us $\neg G$. Let $\mathcal{M}$ be the interpretation that behaves like $\mathcal{K}$ if G and like ID if $\neg G$. Then we have $(\neg G)^{\mathcal{M}}$, without assumptions. Similarly, suppose $\neg G$. Then we find $\Box^*_T G$. Since we have $\Diamond^*_T\top$, it follows that $\Diamond^*_T G$. This gives us an interpretation $\mathcal{N}$ with $G^{\mathcal{N}}$. Reasoning as before we can produce an uncoditional interpretation $\mathcal{P}$ with $G^{\mathcal{P}}$.

4. Suppose $\mathcal{K} : T \rhd U$. We call $\mathcal{K}$ *restricted* (in T) if it admits a 'truth-predicate' K in T, i.e. a predicate K such that $T \vdash K(\underline{\#A}) \leftrightarrow A^{\mathcal{K}}$ for all sentences A of $\mathcal{L}_U$.[13] We write: $\mathcal{K}, K : T \rhd_{\mathsf{res}} U$, etcetera. Suppose $\mathcal{M}, M : T \rhd_{\mathsf{res}} T$. Then the *liarsentence* of M is an oreysentence of T. We leave the amusing verification to the reader.

5. Let ZF^-, be ZF minus the axiom of Foundation Fo. Then Fo is an oreysentence of ZF^-.

6. The Continuum Hypothesis CH is an oreysentence of ZF.

7. The ordinary rossersentence R_{ZF} of ZF is an oreysentence of GB. However, neither $\top \rhd_{\mathsf{ZF}} R$, nor $\top \rhd_{\mathsf{ZF}} \neg R$.

[13] To give a fuller description of what such a predicate would involve is both tedious and laborious. Let's say it is beyond the scope of this paper.

It is an open question whether every sequential theory T that contains enough arithemetic has an oreysentence. All theories that occur in the literature as 'natural' theories do have oreysentences.

4 The language of interpretability logic

The language of interpretability logic, $\mathcal{L}_{\mathsf{int}}$, is the language of modal propositional logic extended with a binary modal operator $\rhd$. We will write $\phi \equiv \psi$ as an abbreviation of $(\phi \rhd \psi \wedge \psi \rhd \phi)$.

Let U be any given theory (in the sense of subsection 3.4). An interpretation $(.)^*$ of $\mathcal{L}_{\mathsf{int}}$ into U maps the atoms on sentences of $\mathcal{L}_U$, commutes with the propositional connectives and satisfies:

$$(\Box\phi)^* := \Box_U \phi^* \text{ and } (\phi \rhd \psi)^* := \phi^* \rhd_U \psi^*.$$

We study the interpretability principles valid in theories U, i.e. we ask ourselves for which χ do we have: for all $(.)^*$: $U \vdash \chi^*$. We will call the set of these principles: $\mathsf{IL}(U)$.

5 The logic IL

We introduce our basic modal logic IL . The principles of our logic are arithmetically sound for a wide class of theories and for various interpretations of its main connective $\rhd$. The theory is not arithmetically complete for any known interpretation. The motivation for studying this specific set of axioms comes from its *modal* simplicity and elegance. The aim of this section is to introduce the logic and to convince the reader of its richness and beauty.

5.1 The logic introduced

IL is the smallest logic in $\mathcal{L}_{\mathsf{int}}$ containing the tautologies of propositional logic, closed under modus ponens and the following rules. (A principle is just a rule with empty antecedent.)

L1 $\vdash \phi \Rightarrow \vdash \Box\phi$

L2 $\vdash \Box(\phi \to \psi) \to (\Box\phi \to \Box\psi)$

L3 $\vdash \Box\phi \to \Box\Box\phi$

L4 $\vdash \Box(\Box\phi \to \phi) \to \Box\phi$

J1 $\vdash \Box(\phi \to \psi) \to \phi \rhd \psi$

J2 $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

J3 $\vdash (\phi \rhd \chi \wedge \psi \rhd \chi) \rightarrow (\phi \vee \psi) \rhd \chi$

J4 $\vdash \phi \rhd \psi \rightarrow (\Diamond\phi \rightarrow \Diamond\psi)$

J5 $\vdash \Diamond\phi \rhd \phi$

IL is valid in all reasonable theories U (i.e. sequential theories containing enough arithmetic, see subsection 3.4). L1-4 are the principles of Löb's Logic. The validity of J1 is witnessed by the identity interpretation ID. J2 reflects the fact that interpretations can be composed. J3 is valid, since, given any two interpretations $\mathcal{K}$ and $\mathcal{M}$ and any sentence A, we can construct an interpretation $\mathcal{N}$ that behaves like $\mathcal{K}$ if A and like $\mathcal{M}$ if $\neg A$. J4 tells us that relative interpretability implies relative consistency. Finally J5 is the 'Interpretation Existence Lemma'. It is valid because a form of Henkin's model construction can be formalized in weak arithmetics like $I\Delta_0 + \Omega_1$. This construction is discussed in section 3.3, example 8. Note that we do *not* have $\vdash (\phi \rhd \psi \wedge \phi \rhd \chi) \rightarrow \phi \rhd (\psi \wedge \chi)$. This principle is invalid as is illustrated by the existence of oreysentences.

De Jongh and Visser prove that IL has *unique and explicit fixed points*. See [13]. No characterization of the *closed fragment* of IL has been given. (The counterexample to the finite model property for simplified models in subsection 5.4 illustrates the richness of the closed fragment.) It is unknown whether IL satisfies *interpolation*. De Jongh and Veltman prove a *modal completeness theorem* w.r.t. Veltman models. See [12]. Veltman models are explained in subsection 5.4.

5.2 Consequences of IL

In our representations of reasoning we always suppress the propositional part. We may reason as follows. We will first prove the principle K3.

K3 $\vdash \Box\phi \leftrightarrow \neg\phi \rhd \bot$

This principle shows that we we have the option to treat the $\Box$ as a defined symbol. Some reformulations in this spirit will be discussed in subsection 5.3.

$$\begin{array}{lcll} \vdash \Box\phi & \rightarrow & \Box(\neg\phi \rightarrow \bot) & \mathsf{L1, L2} \\ & \rightarrow & \neg\phi \rhd \bot & \mathsf{J1} \\ & \rightarrow & \Diamond\neg\phi \rightarrow \Diamond\bot & \mathsf{J4} \\ & \rightarrow & \Box\phi & \mathsf{L1, L2} \end{array}$$

Next we show:

K1 $\vdash \phi \equiv (\phi \vee \Diamond\phi)$

We have:

$$\begin{array}{llcc}
\vdash & \Box(\phi \to \phi) \wedge \Diamond\phi \rhd \phi & \Rightarrow & \mathsf{L1}, \mathsf{J5} \\
\vdash & \phi \rhd \phi \wedge \Diamond\phi \rhd \phi & \Rightarrow & \mathsf{J1} \\
\vdash & (\phi \vee \Diamond\phi) \rhd \phi & & \mathsf{J3} \\
 & & & \\
\vdash & \Box(\phi \to (\phi \vee \Diamond\phi)) & \Rightarrow & \mathsf{L1} \\
\vdash & \phi \rhd (\phi \vee \Diamond\phi) & & \mathsf{J1}
\end{array}$$

We are now ready and set to prove K2.

K2 $\phi \equiv (\phi \wedge \Box\neg\phi)$.

First we reason in Löb's Logic, as follows.

$$\begin{array}{lcll}
\vdash ((\phi \wedge \Box\neg\phi) \vee \Diamond(\phi \wedge \Box\neg\phi)) & \leftrightarrow & ((\phi \wedge \Box\neg\phi) \vee \neg\Box(\Box\neg\phi \to \neg\phi)) & \mathsf{L1}, \mathsf{L2} \\
 & \leftrightarrow & (\phi \wedge \Box\neg\phi) \vee \neg\Box\neg\phi) & \mathsf{L4} \\
 & \leftrightarrow & (\phi \vee \Diamond\phi) &
\end{array}$$

By J1,J2, we may conclude:

$$* \vdash ((\phi \wedge \Box\neg\phi) \vee \Diamond(\phi \wedge \Box\neg\phi)) \equiv (\phi \vee \Diamond\phi)$$

Thus, we find:

$$\begin{array}{lcll}
\vdash (\phi \wedge \Box\neg\phi) & \equiv & ((\phi \wedge \Box\neg\phi) \vee \Diamond(\phi \wedge \Box\neg\phi)) & \mathsf{K1} \\
 & \equiv & (\phi \vee \Diamond\phi) & * \\
 & \equiv & \phi & \mathsf{K1}
\end{array}$$

We give a slightly 'strengthened' version of K2.

K4 $(\phi \wedge \Box\chi) \equiv (\phi \wedge \Box\chi \wedge \Box\neg\phi)$.

We leave the proof of K4 to the reader. An immediate consequence of K2 is the familiar fact that $\vdash \top \rhd \Box\bot$, which was proved first in Feferman's classical 'Arithmetization of Metamathematics' ([20]). Note that this last principle is one possible 'interpretation' version of the Second Incompleteness Theorem. Any theory interprets itself plus its own inconsistency and, hence, cannot prove its own consistency. In the same vein the existence of an oreysentence would be an interpretation version of Rosser's Theorem. In locally essentially reflexive theories, however, oreysentences must have higher complexity than rossersentences.

5.3 Alternative language, alternative axiomatization

In IL we have $\vdash \Box\phi \leftrightarrow \neg\phi \rhd \bot$. So we have the possibility to eliminate $\Box$ from the language. Moreover, some axioms are superfluous, so we can give a more efficient axiomatization. Eliminating the $\Box$ can be often convenient, e.g. in proving modal completeness theorems. We present and verify one alternative.

Let $\mathcal{L}^{-}_{\mathsf{int}}$ be the language of interpretability logic without $\Box$. We will treat $\Box\phi$ as an abbreviation of $\neg\phi \rhd \bot$. Thus there is an obvious translation of $\mathcal{L}_{\mathsf{int}}$ to $\mathcal{L}^{-}_{\mathsf{int}}$. To spare ourselves the anxiety of carrying scholastic distinctions along, I will in the following simply confuse the two languages. IL can be axiomatized as the smallest logic in $\mathcal{L}_{\mathsf{int}}$ containing the tautologies of propositional logic, closed under modus ponens and the following rules.

I1 $\vdash \phi \to \psi \Rightarrow \vdash \phi \rhd \psi$

I2 $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

I3 $\vdash (\phi \rhd \chi \wedge \psi \rhd \chi) \to (\phi \vee \psi) \rhd \chi$

I4 $\vdash \phi \rhd (\phi \wedge (\phi \rhd \bot))$

Proof

It is easy to see that these principles can be derived from IL. I4 is just a variant of K2. We prove the usual principles from the alternative ones. Clearly, J2=I2 and J3=I3. Using I1 and I2 and propositional logic one easily verifies the substitution principle sub: $\vdash \phi \leftrightarrow \psi \Rightarrow \vdash \chi[p := \phi] \leftrightarrow \chi[p := \chi]$. Thus we certainly do not have to worry about replacing subformulas by their equivalents in propositional logic. E.g., $\phi \rhd \bot$ can be interchanged with $\Box\neg\phi$. In the verifications below we will mostly suppress mention of the use of I1, I2, sub and propositional logic.

L1 Suppose $\vdash \phi$, then $\vdash \neg\phi \to \bot$ and, hence, $\vdash \Box\phi$.

J1 Reason inside $\vdash$. Suppose $\Box(\phi \to \psi)$. We have $(\phi \wedge \neg\psi) \rhd \bot$ and, a fortiori, $(\phi \wedge \neg\psi) \rhd \psi$. Also, $\psi \rhd \psi$, hence, by I3, $((\phi \wedge \neg\psi) \vee \psi) \rhd \psi$. So, finally, $\phi \rhd \psi$.

L2 Reason inside $\vdash$. Suppose $\Box(\phi \to \psi)$. Then $\Box(\neg\psi \to \neg\phi)$, so, by J1, $\neg\psi \rhd \neg\phi$. Suppose $\Box\phi$, i.e., $\neg\phi \rhd \bot$. By I2, $\neg\psi \rhd \bot$, i.e., $\Box\psi$.

J4 Reason inside $\vdash$. Suppose $\Box(\phi \to \psi)$. Then, by J1, $\phi \rhd \psi$. We reason by contraposition. Suppose $\Box\neg\psi$ and, thus, $\psi \rhd \bot$. By I2, $\phi \rhd \bot$ and, thus, $\Box\neg\phi$.

L4 Reason inside $\vdash$. By I4, $\neg\phi \rhd (\neg\phi \wedge \Box\phi)$. So, by J4, $\Diamond\neg\phi \to \Diamond(\neg\phi \wedge \Box\phi)$. Hence, $\Box(\Box\phi \to \phi) \to \Box\phi$.

J5 Reason inside $\vdash$. We have $\Diamond\phi \rhd (\phi \vee \Diamond\phi)$. Moreover, by I4,

$$(\phi \vee \Diamond\phi) \rhd ((\phi \vee \Diamond\phi) \wedge (\phi \vee \Diamond\phi) \rhd \bot).$$

Hence, $(\phi \vee \Diamond\phi) \rhd ((\phi \vee \Diamond\phi) \wedge \Box\neg\phi)$, and so $\Diamond\phi \rhd \phi$.

L3 As is well known, L3 follows from L1,L2,L4. We give an alternative derivation. Reason inside $\vdash$. J5 gives us: $\Diamond\neg\phi \rhd \neg\phi$. Hence, by J4, $\Diamond\Diamond\neg\phi \to \Diamond\neg\phi$. Ergo: $\Box\phi \to \Box\Box\phi$.

❑

A variation of our theme is to study the contraposed versions of interpretability, Π_1-conservativity and the like. We write $A \rhd_{\mathsf{c\text{-}X}} B$ for $\neg B \rhd_{\mathsf{X}} \neg A$. E.g:

- $A \rhd_{\mathsf{c\text{-}int}} B \Leftrightarrow \exists\mathcal{K}\ \Box(A^{\mathcal{K}} \to B)$,
- $A \rhd_{\mathsf{c}\text{-}\Pi_1\text{-}\mathsf{cons}} B \Leftrightarrow \forall S{\in}\Sigma_1\ (\Box(S \to A) \to \Box(S \to B))$.

I will call contraposed Π_1-conservativity: *Σ_1-preservativity.* As we will see in subsection 10.4 Σ_1-preservativity *in the precise form given above* is a more interesting notion than Π_1-conservativity as soon as we turn to constructive logic. For sequential locally essentially reflexive theories that satisfy full Σ-completeness, both notions have an Orey-Hájek characterization:

- $A \rhd_{\mathsf{c\text{-}int}} B \Leftrightarrow A \rhd_{\Sigma_1\text{-}\mathsf{pres}} B \Leftrightarrow \forall n\ \Box(\Box_n A \to B)$.

The principles of contraposed IL are as follows.

ι1 $\vdash \phi \to \psi \Rightarrow \vdash \phi \rhd \psi$

ι2 $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

ι3 $\vdash (\chi \rhd \phi \wedge \chi \rhd \psi) \to \chi \rhd (\phi{\wedge}\psi)$

ι4 $\vdash (\Box\phi \to \phi) \rhd \phi$

Here $\Box\phi$ abbreviates $\top \rhd \phi$. If we rewrite $\phi \rhd \psi$ as $[\phi]\psi$, we see that the $[\phi]$'s are necessity operators of a normal modal logic. To get some feeling for the logic the reader might amuse him/herself by deriving contraposed J5, i.e. $\vdash \phi \rhd \Box\phi$, from scratch. Note that in contraposed form oreysentences appear as sentences O with the property: $O \rhd \bot$ and $\neg O \rhd \bot$.

5.4 Semantics

A Kripke semantics for IL was discovered by Frank Veltman. An IL-frame (or Veltman frame) is a tuple $\langle K, R, S\rangle$, where:

- K is a non-empty set
- R is a transitive, upwards wellfounded relation on K
- S is a ternary relation on K satisfying:
  - $yS_xz \Rightarrow xRy$ and xRz

- $xRyRz \Rightarrow yS_xz$
- S_x is transitive and reflexive on $\{y \mid xRy\}$

A forcing relation $\Vdash$ on an IL-frame satisfies the usual clauses, where R is the accessibility relation for $\Box$, plus:

- $x \Vdash \phi \rhd \psi \Leftrightarrow \forall y((xRy \wedge y \Vdash \phi) \to \exists z(yS_xz \wedge z \Vdash \psi))$

This makes $\rhd$ into a sort of *might-conditional.*

An IL-model or Veltman model is a structure $\langle K, R, S, \Vdash\rangle$, where $\langle K, R, S\rangle$ is an IL-frame and $\Vdash$ is a forcing relation on $\langle K, R, S\rangle$. Veltman & de Jongh show: IL is sound & complete for finite IL-models. See [12].

For many logics extending IL one can get rid of the subscript in the S relation. These models are simplified Veltman models, sometimes called Visser models. Their full definition is as follows.

- K is a non-empty set
- R is a transitive, upwards wellfounded relation on K
- S is a transitive, reflexive relation on K satisfying $R \subseteq S$

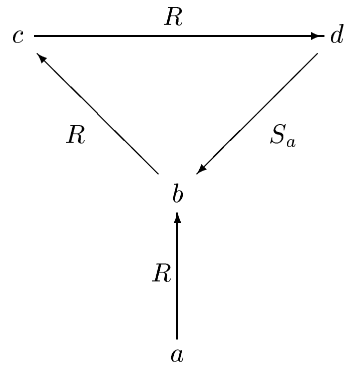
We can recover the S_x's in the new setting by taking:

$$yS_xz \;:\Leftrightarrow\; xRy,\ xRz \text{ and } ySz.$$

The clauses for forcing are as before, now using the *defined* S_x's. In [53] it is shown that we can unravel each Veltman model to a bisimulating simplified one. (The relevant notion of bisimulation is specified below.) So we have also completeness for IL in simplified models. The following example shows that we lose the finite model property for IL if we work with simplified models. Consider the formula:

$$\phi \;:=\; \Diamond\top \wedge \top \rhd \Diamond\Diamond\top \wedge \Box(\Diamond\top \to \neg(\top \rhd \Diamond\top))$$

Here is a Veltman model satisfying this formula.



We employ the obvious convention that the model intended is the closure under the closure rules of IL-models. Thus we will have: cS_bd, dS_ac, but not dS_bc. It is easy to see that $a \Vdash \phi$. Consider an arbitrary simplified model $\mathcal{K}$ with a node a with $a \Vdash \phi$. We will show that $\mathcal{K}$ is not finite.

Proof

Since $a \Vdash \Diamond\top$ and $a \Vdash \top \rhd \Diamond\Diamond\top$, we can find a b with $aRb \Vdash \Diamond\Diamond\top$. A fortiori, $b \Vdash \Diamond\top$. Because $a \Vdash \Box(\Diamond\top \to \neg(\top \rhd \Diamond\top))$, there is a d' with bRd' and, for *no* c, bRc, $d'Sc$ and $c \Vdash \Diamond\top$. Let d be a top node of our model such that $d = d'$ or $d'Rd$. Clearly $d \Vdash \Box\bot$. It is easy to see that, by the closure conditions of Veltman models, bRd and, for *no* c, bRc, dSc and $c \Vdash \Diamond\top$. We claim that not dSb. Since $b \Vdash \Diamond\Diamond\top$, there is a c with $bRc \Vdash \Diamond\top$. If we had dSb, it would follow that $dSbSc$, and, thus, dSc. Quod non. Since $a \Vdash \top \rhd \Diamond\Diamond\top$, there is a b' with aRb', dSb' and $b' \Vdash \Diamond\Diamond\top$.

Thus, we can construct a chain of nodes, $b_1, d_1, b_2, d_2 \ldots$, such that aRb_iRd_i, d_iSb_{i+1}, not d_iSb_i, and $d_i \Vdash \Box\bot$. Note that if x occurs in the chain before y, then xSy. Now assume that our model is finite. It follows that some node e must occur twice in the chain. Since nodes of the b-type are necessarily distinct from nodes of the d-type, it follows that we can construct a cycle containing some pair b_i, d_i. Clearly any two nodes on the cycle will be S-related, and hence d_iSb_i. A contradiction. ❑

The study of models for IL leads to a strenghtened notion of bisimulation. Let two models $\mathcal{K}$ and $\mathcal{M}$ be given. A relation $\mathcal{Z}$ between the nodes of our models is a *bisimulation* if it satisfies the following conditions.

at $k\mathcal{Z}m \Rightarrow (k \Vdash p \Leftrightarrow m \Vdash p)$, for all atoms p,

zig If kRk' and $k\mathcal{Z}m$, then there is a m' with mRm' and $k'\mathcal{Z}m'$ and, for all m'' with $m'S_mm''$, there is a k'' with $k''\mathcal{Z}m''$ and $k'S_kk''$,

zag If mRm' and $k\mathcal{Z}m$, then there is a k' with kRk' and $k'\mathcal{Z}m'$ and, for all k'' with $k'S_kk''$, there is a m'' with $k''\mathcal{Z}m''$ and $m'S_mm''$.

Since we can associate a model to a simplified model in the evident way, it makes sense to speak about bisimulations between models and simplified models and between simplified models and simplified models. We will see bisimulations in action is subsection 8.2. We end this subsection with a picture of the zig-clause.

$$\begin{tikzcd}
k' \arrow[r, dashed, "S_k"] & k'' & & m' \arrow[r, "S_m"] & m'' \\
k \arrow[u, "R"] & & & m \arrow[u, dashed, "R"] &
\end{tikzcd}$$

6 F, W, W* and other principles

In this section we discuss some further principles in interpretability logic. All the principles discussed in this section are valid in all reasonable arithmetical theories.

$$\begin{array}{rl}
\mathsf{F} & \vdash \Diamond\phi \to \neg(\phi \rhd \Diamond\phi)\\[1ex]
\mathsf{KW2} & \vdash \phi \rhd \Diamond\psi \to \psi \rhd (\psi\wedge\neg\phi)\\[1ex]
\mathsf{W} & \vdash \phi \rhd \psi \to \phi \rhd (\psi\wedge\Box\neg\phi)\\[1ex]
\mathsf{KW3} & \vdash \phi \rhd (\psi\vee\Diamond\phi) \to \phi \rhd \psi\\[1ex]
\mathsf{KW4} & \vdash \phi \rhd \psi \to \psi \rhd (\psi\wedge\Box\neg\phi)\\[1ex]
\mathsf{M}_0 & \vdash \phi \rhd \psi \to (\Diamond\phi\wedge\Box\chi) \rhd (\psi\wedge\Box\chi)\\[1ex]
\mathrm{W}^* & \vdash \phi \rhd \psi \to (\psi\wedge\Box\chi) \rhd (\psi\wedge\Box\chi\wedge\Box\neg\phi)
\end{array}$$

Before discussing the meaning of these principles, let me describe their interrelations. It is easily seen that W follows from W*. We prove that M_0 follows from W*.

$$\begin{array}{rcll}
\vdash \phi \rhd \psi & \to & \phi \rhd (\psi\vee\Diamond\phi) & \mathsf{J1, J2}\\
 & \to & ((\psi\vee\Diamond\phi)\wedge\Box\chi) \rhd ((\psi\vee\Diamond\phi)\wedge\Box\chi\wedge\Box\neg\phi) & \mathsf{W}^*\\
 & \to & (\Diamond\phi\wedge\Box\chi) \rhd (\psi\wedge\Box\chi) & \mathsf{J1, J2}
\end{array}$$

Conversely, W* follows from W and M_0 together. The argument is due to Dick de Jongh.

$$\begin{array}{rcll}
\vdash \phi \rhd \psi & \to & \phi \rhd (\psi\wedge\Box\neg\phi) & \mathsf{W}\\
 & \to & (\Diamond\phi\wedge\Box\chi) \rhd (\psi\wedge\Box\neg\phi\wedge\Box\chi) & \mathsf{M}_0\\
 & \to & ((\psi\wedge\Diamond\phi\wedge\Box\chi)\vee(\psi\wedge\Box\neg\phi\wedge\Box\chi)) \rhd (\psi\wedge\Box\neg\phi\wedge\Box\chi) & \mathsf{J1, J2, J3}\\
 & \to & (\psi\wedge\Box\chi) \rhd (\psi\wedge\Box\chi\wedge\Box\neg\phi) & \mathsf{J1, J2}
\end{array}$$

W, KW3 and KW4 are interderivable over IL. Most of the arguments are simple. We show that KW4 follows from W. We leave applications of L1,L2 implicit in this proof.

$$\begin{array}{lcl}
\vdash \psi \rhd ((\psi\wedge\Box\neg\phi)\vee\Diamond\phi) & \Rightarrow & \mathsf{J1}\\
\vdash \psi \rhd ((\psi\wedge\Box\neg\phi)\vee\phi) & \Rightarrow & \mathsf{J1, J2, J5, J3}\\
\vdash \phi \rhd \psi \to \psi \rhd (\psi\wedge\Box\neg\phi) & & \mathsf{J1, J2, W, J3}
\end{array}$$

Here is the derivation of KW2 from W.

$$\begin{array}{rcll}
\vdash \phi \rhd \Diamond\psi & \to & \phi \rhd (\Diamond\psi\wedge\Box\neg\phi) & \mathsf{W}\\
 & \to & \phi \rhd \Diamond(\psi\wedge\neg\phi) & \mathsf{L1, L2, J1, J2}\\
 & \to & \phi \rhd (\psi\wedge\neg\phi) & \mathsf{J5, J2}\\
 & \to & ((\psi\wedge\neg\phi)\vee\phi) \rhd (\psi\wedge\neg\phi) & \mathsf{J1, J2, J3}\\
 & \to & \psi \rhd (\psi\wedge\neg\phi) & \mathsf{J1, J2}
\end{array}$$

F easily follows from KW2.

Our principle KW2 is interderivable with Švejdar's KW1^0. Vítěslav Švejdar in his [46] established that KW2 (in his paper KW1^0) is not derivable from F over IL. He also shows that W is not derivable from KW2. Mladen Vuković shows that W is not derivable from M_0. See his [61]. It is easy to see W^* (and, hence, M_0) does not follow from W —these principles correspond to different classes of frames. Here is a schema of the dependencies of our principles over IL.

$\vdash$	F	KW2	W	KW3	KW4	M_0	W+M_0	W^*
F	+	−	−	−	−	−	−	−
KW2	+	+	−	−	−	−	−	−
W	+	+	+	+	+	−	−	−
KW3	+	+	+	+	+	−	−	−
KW4	+	+	+	+	+	−	−	−
M_0	?	?	−	−	−	+	−	−
W+M_0	+	+	+	+	+	+	+	+
W^*	+	+	+	+	+	+	+	+

F, KW2 and W all characterize the same class of Veltman frames: the class of frames such that for each x, $R \circ S_x$ is upwards wellfounded. Such frames are called ILW-frames. Thus ILF and IL(KW2) are incomplete w.r.t. their characteristic classes. De Jongh and Veltman prove the completeness theorem for ILW w.r.t. finite ILW-models. Unfortunately, their proof —which is rumoured to be very beautiful— was never published. Simplified ILW-frames, are, as expected, simplified IL-frames with the extra property that $R \circ S$ is upwards wellfounded. In [53] it is shown that for every ILW-model there is a bisimilar simplified ILW-model. Moreover the construction preserves finiteness. Thus, we have completeness for ILW in finite, simplified ILW-models.

The characteristic class of both ILM_0 and ILW^* is the class of frames satisfying: $xRyRzS_xuRv \Rightarrow yRv$. It follows that ILM_0 is incomplete w.r.t. its characteristic class. No modal completeness theorem is know for ILW^*.

The *closed fragment* of ILF (and, ipso facto, of all extensions of ILF) has been characterized by Hájek and Švejdar: it is the same as the one of Löb's logic. See [26]. It is unknown whether ILF or ILW satisfies *interpolation*. We will see in subsection 8.2 that ILM_0 and ILW^* do not satisfy interpolation.

F can be viewed as an interpretability version of the second incompleteness theorem. ILW was conjectured to be the interpretability logic of all reasonable arithmetical theories. This conjecture was refuted in [56]. The current conjecture is that ILW^* is the interpretability logic of all reasonable arithmetical theories.

An immediate consequence of KW2 is the Contraposition Principle:

KW1 $\vdash \phi \rhd \Diamond\top \to \top \rhd \neg\phi$

We give an application of this principle. Paris & Wilkie show that

$$\mathsf{EXP} \rhd_{I\Delta_0+\Omega_1} \Diamond_{I\Delta_0+\Omega_1}\top$$

(see [35]), ergo by KW1: $\top \rhd_{I\Delta_0+\Omega_1} \neg\mathsf{EXP}$, i.o.w.:

$$(I\Delta_0+\Omega_1) \rhd (I\Delta_0+\Omega_1+\neg\mathsf{EXP}).$$

This fact was originally shown by Solovay employing vastly different means.

7 The principle P

P is the Persistence Principle:

P $\vdash \phi \rhd \psi \to \Box(\phi \rhd \psi)$

P characterizes IL-frames with the following property:

$$yRzS_xu \Rightarrow (yRu \wedge zS_yu).$$

We call such frames ILP-frames. De Jongh & Veltman show the completeness of ILP w.r.t. (finite) ILP-models. See their [12]. Simplified ILP-frames will be IL-frames with the extra property: $yRzSu \Rightarrow yRu$. In [53], it is shown that any ILP-model bisimulates with a simplified ILP-model. Moreover, this construction preserves finiteness. Thus, we have completeness for ILP in finite, simplified ILP-models. It is unknown whether ILP satisfies interpolation!

P is valid for interpretations in finitely axiomatized arithmetical theories extending, say, $I\Delta_0+\Omega_1$. To see the arithmetical validity, reason as follows. Let C be the single axiom of T. Suppose $A \rhd_T B$. This means that, for some $\mathcal{K}$, $\Box_T(A \to (B^{\mathcal{K}} \wedge C^{\mathcal{K}}))$. It follows that $\Box_T\Box_T(A \to (B^{\mathcal{K}} \wedge C^{\mathcal{K}}))$, i.o.w. $\Box_T(A \rhd_T B)$. ILP is complete for interpretations in finitely axiomatized sequential theories with designated natural numbers satisfying $I\Delta_0+\mathsf{SUPEXP}$ that do not prove their iterated inconsistency for any finite number of iterations. Examples of such theories are: $I\Delta_0+\mathsf{SUPEXP}$, $I\Sigma_n$ for $n=1,2,\ldots$, ACA_0 and GB. The proof of arithmetical completeness is given in [55]. For a somewhat different proof, not using simplified Veltman models, see [62]. For a proof of modal and arithmetical completeness for the closely related ILP^{ω}, see [15].

Since ILP is arithmetically complete for *some* class of reasonable theories, it must, a fortiori, imply W*, which is valid in *all* reasonable theories. We verify directly that ILP extends ILW*.

$$\begin{array}{rcll}
\vdash \phi \rhd \psi & \to & \Box(\phi \rhd \psi) & \mathsf{P} \\
& \to & \Box(\Diamond\phi \to \Diamond\psi) & \mathsf{J4} \\
& \to & \Box((\psi\wedge\Box\chi\wedge\Box\neg\psi) \to (\psi\wedge\Box\chi\wedge\Box\neg\phi)) & \mathsf{L1}, \mathsf{L2} \\
& \to & (\psi\wedge\Box\chi\wedge\Box\neg\psi) \rhd (\psi\wedge\Box\chi\wedge\Box\neg\phi) & \mathsf{J1} \\
& \to & (\psi\wedge\Box\chi) \rhd (\psi\wedge\Box\chi\wedge\Box\neg\phi) & \mathsf{K4}, \mathsf{J2}
\end{array}$$

Note that in fact we proved:

$$\mathsf{IL} \vdash \Box(\phi \rhd \psi) \to (\psi \wedge \Box\chi) \rhd (\psi \wedge \Box\chi \wedge \Box\neg\phi).$$

8 Montagna's Principle M

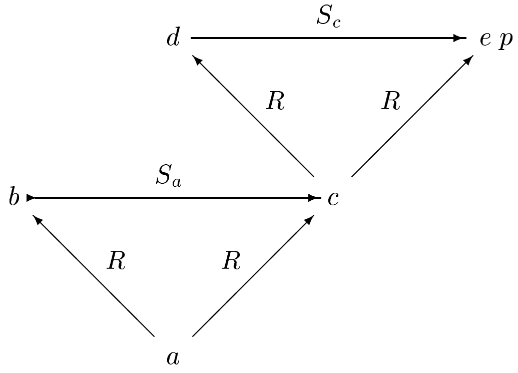
We exhibit a principle which is arithmetically valid in all verifiably locally essentially reflexive theories, Montagna's Principle M:

$\mathrm{M} \vdash \phi \rhd \psi \to (\phi \wedge \Box\chi) \rhd (\psi \wedge \Box\chi)$

M was known before to Lindström and to Švejdar (even if not in 'modal guise'). M characterizes IL-frames with the following property: $yS_xzRu \Rightarrow yRu$. We call such frames ILM-frames. De Jongh & Veltman show that ILM is complete w.r.t. finite ILM-models. See [12], or, alternatively, [5]. A simplified ILM-frame is a simplified IL-frame with the property that $ySzRu \Rightarrow yRu$. In [53] it is shown that every ILM-model bisimulates with a simplified ILM-model. The proof is also given in the more accessible [5]. Thus we have completeness w.r.t. simplified ILM-models. However for simplified ILM-models we do not have the finite model property, as is illustrated by the following formula.

$$\Diamond\Diamond\top \wedge \Diamond\top \rhd (\Diamond\top \wedge \top \rhd p) \wedge \Diamond\top \rhd (\Diamond\top \wedge \neg(\top \rhd p))$$

Here is an ordinary Veltman model satisfying this formula (in a).



Suppose there would be a finite simplified ILM-model satisfying our formula. Then there would be a sequence of nodes, $b_1, c_1, b_2, c_2, \ldots$, such that b_i and c_i are R-above a, b_iSc_i, $b_i \Vdash (\Diamond\top \wedge \top \rhd p)$ and $c_i \Vdash (\Diamond\top \wedge \neg(\top \rhd p))$. Since our model is finite there must be b_i, c_i with c_iSb_i. It is however easy to see that in a simplified ILM model any two nodes that are on an S-cycle and that force the same atoms must force the same formulas. So we have a contradiction.

M is arithmetically valid in all verifiably locally essentially reflexive theories T extending EA ($= I\Delta_0 + \mathsf{EXP}$). To see this, fix such a theory T. Remember that, by fact 3.2, in T we have that interpretability over T extensionally conincides with Π_1-conservativity over T. So it is sufficient to prove M for Π_1-conservativity over T. Reason in T. Suppose $A \rhd_{\Pi_1\text{-}\mathsf{cons},T} B$. Let S be any Σ_1-sentence and P be any Π_1-sentence. Suppose $\Box_T((B{\wedge}S) \to P)$. Then, $\Box_T(B \to (\neg S{\vee}P))$, hence, since $(\neg S{\vee}P)$ is Π_1, $\Box_T(A \to (\neg S{\vee}P))$. Ergo, $\Box_T((A{\wedge}S) \to P)$. Noting that we may replace S by $\Box_T C$, we are done.

For extensions of PA in the language of arithmetic, we can see M immediately by reflecting upon the following well-known characterization.

$$U \rhd V \Leftrightarrow \text{ all } \mathcal{M} \models U \text{ have end-extension } \mathcal{N} \models V.$$

Alessandro Berarducci and Volodya Shavrukov have shown (independently) that ILM is complete for arithmetical interpretations in extensions T of PA in the language of PA that are Σ^0_1-sound, or, more precisely, that do not prove one of their own iterated inconsistency statements, $\Box^n_T\bot$. See their papers [5] and [39]. For a proof not using simplified Veltman models, see [62]. Hájek & Montagna show that ILM is complete for arithmetical interpretations for Π_1-conservativity in extensions of $I\Sigma_1$ that do not prove their own iterated inconsistency statements. See their paper [23]. For a proof that avoids simplified Veltman models, see [19].

Since in verifiably locally essentially reflexive theories T extending EA, interpretability and Π_1-conservativity over T provably coincide (see fact 3.2), the result by Hájek and Montagna tells us that ILM is complete for arithmetical interpretations in verifiably locally essentially reflexive theories T extending $I\Sigma_1$ that do not prove one of their own iterated inconsistency statements.[14] The weakest such theory is $I\Sigma_1 + \mathsf{Rfn}_{I\Sigma_1}$, where $\mathsf{Rfn}_{I\Sigma_1}$ is the local reflection principle for $I\Sigma_1$. This theory is certainly below PA, since, by an observation of Feferman, it is a subtheory of $I\Sigma_1$ plus the set of all true Π_1-sentences.

Clearly, since ILM is complete for some class of interpretations in reasonable arithmetical theories, W^* must be derivable in ILM. We show how to do this. First we derive W. Reason in ILM. Suppose we have $\phi \rhd \psi$. By M we may conclude: $(\phi{\wedge}\Box\neg\phi) \rhd (\psi{\wedge}\Box\neg\phi)$. By $\mathsf{K2}$, we have $\phi \rhd (\phi{\wedge}\Box\neg\phi)$. Hence, using $\mathsf{J2}$, $\phi \rhd (\psi{\wedge}\Box\neg\phi)$. Using W, we may derive $\mathsf{KW4}$. So we find, from $\phi \rhd \psi$, $\psi \rhd (\psi{\wedge}\Box\neg\phi)$. Applying M again, we have $(\psi{\wedge}\Box\chi) \rhd (\psi{\wedge}\Box\chi{\wedge}\Box\neg\phi)$.

8.1 Consequences of M

Two consequences of M are:

$\mathsf{KM1}$ $\vdash \phi \rhd \Diamond\psi \to \Box(\phi \to \Diamond\psi)$

[14] In the proof of the theorem a primitive recursive function is introduced of which it is shown that it is eventually weakly decreasing. Then one needs to infer that it is eventually constant. It is precisely at this point that Σ_1-induction is needed.

$\mathsf{KM2} \vdash \phi \rhd \psi \to (\Box(\psi \to \Diamond\chi) \to \Box(\phi \to \Diamond\chi))$

Clearly, these principles show us what is 'visible' of the Π_1-conservativety of essentially reflexive theories over theories interpreted in them. It is easily seen that these principles are interderivable over IL. Both $\mathsf{KM1}$ and $\mathsf{KM2}$ characterize IL-frames satisfying $yS_xzRu \Rightarrow yRu$. Švejdar shows that neither of them implies M over IL. See his [46].

8.2 Failure of interpolation

Consider any logic $\mathcal{I}$ between ILM_0 and ILM. We show that $\mathcal{I}$ does not satisfy interpolation. The proof is a minor adaptation of Konstantin Ignatiev's unpublished argument that ILM does not satisfy interpolation. Note that it follows that ILW^* does not have interpolation and neither has $\mathsf{IL}(\forall)$, the interpretability logic of all reasonable theories, which is after all firmly in between ILM_0 and ILM.

Proof

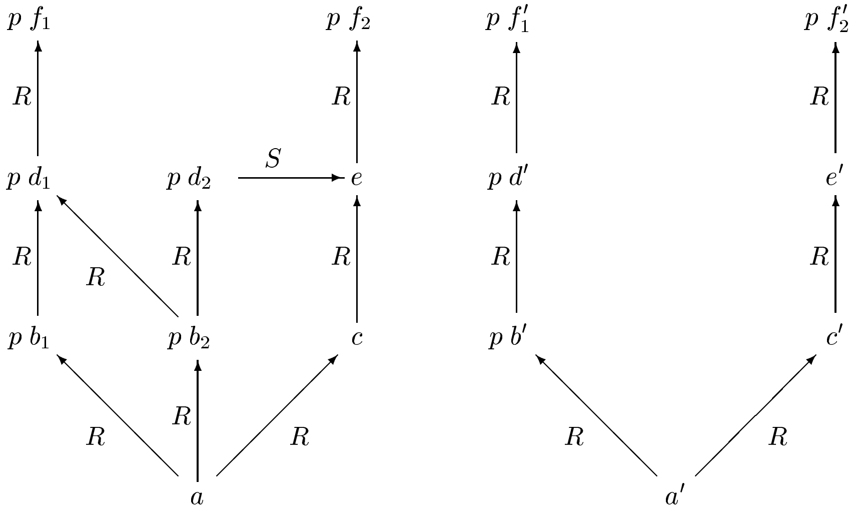
We have:

$$\mathcal{I} \vdash \Box(p \leftrightarrow \Box q) \to (r \rhd t \to (\Diamond r \wedge p) \rhd (t \wedge p)).$$

Suppose, to get a contradiction, that there is a formula $I(p)$ only containing p, such that:

$$\mathcal{I} \vdash \Box(p \leftrightarrow \Box q) \to I(p) \quad \text{and} \quad \mathcal{I} \vdash I(p) \to (r \rhd t \to (\Diamond r \wedge p) \rhd (t \wedge p)).$$

Consider the following two simplified models for ILM.



We employ the usual convention that the intended models are what one gets when one closes off the displayed relations under the appropriate closure rules. So, for example, in the first model we will have $b_2 S f_2$ and $d_2 R f_2$. p is only forced where this is displayed. Let $\mathcal{B}$ be the relation between the nodes of the models that connects f_1, f_2 with f_1', f_2', d_1, d_2 with d', etc. Inspection shows that $\mathcal{B}$ is a bisimulation. The crux is that, for b_2, b', we meet the R-move from b' to d' with an R-move from b_2 to d_1, and that, for a, a', we meet the R-moves from a' to b' and to d' with counter moves from a to b_1, respectively d_1. Since our models are ILM-models, they are, a fortiori, $\mathcal{I}$-models. Extend the second model by stipulating that $d' \Vdash q$ and $f_1' \Vdash q$ (and no other node forces q). We get: $a' \Vdash \Box(p \leftrightarrow \Box q)$. Hence, by assumption: $a' \Vdash I(p)$. By bisimulation, we find $a \Vdash I(p)$. Now extend the first model by stipulating that $d_2 \Vdash r$ and $e \Vdash t$. It follows that $a \Vdash r \rhd t$. Since $a \Vdash I(p)$, we obtain that $a \Vdash (\Diamond r \wedge p) \rhd (t \wedge p))$. Quod non. ❑

8.3 An overview of systems

We end this section with a schema of salient systems and what is and is not known about them. The '??' in the case of the arithmetical completeness question signals that we not only do not have an arithmetical completeness result, but even lack a conjecture on what the appropriate arithmetical semantics should be.

system	kripke comp	arith comp	interpol	clos frag
IL	+	??	?	?
ILF	−	??	?	+
ILW	+	??	?	+
ILW*	?	?	−	+
ILP	+	+	?	+
ILM	+	+	−	+

9 Beyond finite and essentially reflexive

This section provides some remarks concerning the interpretability logics of theories which are not locally essentially reflexive extensions of $I\Sigma_1$ or finitely axiomatized, sequential extensions of $I\Delta_0 + \mathsf{SUPEXP}$.

> WARNING: To make sense of interpretability in weak theories, which almost always lack Σ_1^0-collection, we have to employ the notion of smooth interpretability. See [56].

9.1 Weak theories

We know that the *provability logics* of $I\Delta_0+\Omega_1$ and S^1_2 both extend Löb's Logic. It is a great open question whether these logics are equal to Löb's Logic. (See [48] and [6].) In the light of this great open problem the question what the *interpretability logics* of $I\Delta_0+\Omega_1$ and S^1_2 are, seems to be definitely immodest. However, you never know. Sometimes it is easier to answer a seemingly more difficult question.

The interpretability logics of $I\Delta_0+\Omega_1$ and S^1_2 are certainly *totaliter aliter* compared to the logics we know. Let me just mention two valid principles of both. We will sketch the arguments for their arithmetical validity in $I\Delta_0+\Omega_1$ in the footnotes.

$\wedge\Pi \vdash (\phi \rhd \Diamond\psi \wedge \phi \rhd \Diamond\chi) \to \phi \rhd (\Diamond\psi \wedge \Diamond\chi)$[15]

$\mathsf{P}\Pi \vdash \phi \rhd \Diamond\psi \to \Box(\phi \rhd \Diamond\psi)$[16]

It is easy to see that the first principle is not provable in ILP, and, hence, not generally arithmetically valid. The second principle is in the intersection of ILP and ILM. It is not provable in ILW^*. One can show that it is not in $\mathsf{IL}(\mathsf{PRA})$. See subsection 9.3.

9.2 $I\Delta_0+\mathsf{EXP}$

In [55] it was shown that the following semantics is sound and complete for $\mathsf{IL}(I\Delta_0+\mathsf{EXP})$. Our models are finite strict partial orders, with accessibility relation, say, R. The clauses for atoms and propositional atoms are as usual. The accessibility relation for $\Box$ is $R\circ R$. The clause for $\rhd$ is:

$$k \Vdash \phi \rhd \psi :\Leftrightarrow \forall m\, ((kRm \wedge \exists n\, (mRn \wedge n \Vdash \phi)) \Rightarrow \exists p\, (mRp \wedge p \Vdash \psi)).$$

Clearly this logic can be viewed as a fragment of Löb's Logic via the following translation $(.)^*$. $(.)^*$ commutes with atoms and with the propositional connectives, $(\Box\phi)^* := \Box\Box\phi^*$ and $(\phi \rhd \psi)^* := \Box(\Diamond\phi^* \to \Diamond\psi^*)$. The problem of axiomatizing this logic in a perspicuous way seems to be remarkably hairy. Marianne Kalsbeek provides a number of principles in her preprint [30].

9.3 PRA and its kin

We study the interpretability logic of PRA and some related theories. Since not much is known of this logic, the discussion of this section will be somewhat

[15] If $I\Delta_0+\Omega_1$ plus the arithmetical interpretation of $\Diamond\psi$ is interpretable, it is interpretable on a definable cut, which is closed under ω_1. Similarly for $I\Delta_0+\Omega_1$ plus the arithmetical interpretation of $\Diamond\chi$. The intersection of the two cuts, will, by downwards persistence of Π_1-sentences, interpret $I\Delta_0+\Omega_1$ plus the arithmetical interpretations of both $\Diamond\psi$ and $\Diamond\chi$.

[16] If $I\Delta_0+\Omega_1$ plus the arithmetical interpretation of $\Diamond\psi$ is interpretable, it is interpretable on a definable cut I closed under ω_1. We can define this cut in such a way that $I\Delta_0+\Omega_1$ verifies the statement that, for all A, $I\Delta_0+\Omega_1 \vdash A \Rightarrow I\Delta_0+\Omega_1 \vdash A^I$.

tentative. We show that the interpretability logic of PRA strictly extends the minimal logic. The following reasoning is due to Lev Beklemishev. Every finite Σ^0_2-axiomatized extension of PRA is reflexive. (See [2].) Smooth interpretability in a reflexive theory has the Orey-Hájek characterization —verifiably in PRA. This tells us that $A \rhd_{\mathsf{PRA}} B$, with $A \in \Sigma^0_2$, is Π^0_2. Moreover, again by Orey-Hájek, for $A \in \Sigma^0_2$, we have:

$$\mathsf{PRA} \vdash A \rhd_{\mathsf{PRA}} B \to (A \wedge \Box_{\mathsf{PRA}} C) \rhd_{\mathsf{PRA}} (B \wedge \Box_{\mathsf{PRA}} C).$$

Now define the following class S_2 of modal formulas:

1. boxed formulas and their negations are in S_2
2. if $\phi \in \mathsf{S}_2$ and ψ arbitrary, then $\neg(\phi \rhd \psi) \in \mathsf{S}_2$
3. S_2 is closed under conjunction and disjunction

It is easily seen that all interpretations of formulas in S_2 are in Σ^0_2. So the following principle will be in IL(PRA):

$\mathsf{B} \vdash \phi \rhd \psi \to (\phi \wedge \Box \chi) \rhd (\psi \wedge \Box \chi)$, for $\phi \in \mathsf{S}_2$

Remember that to use Kripke semantics for non-derivability results, we only need soundness. It is easy to show, by a Kripke model argument, that the following instance of B:

$$\Diamond p \rhd q \to (\Diamond p \wedge \Box r) \rhd (q \wedge \Box r)$$

is not in ILW*. So IL(PRA) is not the minimal interpretability logic.

We provide some 'upperbounds' for IL(PRA). Since M is not derivable in ILP, we can find, by the arithmetical completeness theorem for ILP find arithmetical sentences A, B, C such that:

$$I\Sigma_1 \nvdash A \rhd_{I\Sigma_1} B \to (A \wedge \Box_{I\Sigma_1} C) \rhd_{I\Sigma_1} (B \wedge \Box_{I\Sigma_1} C)$$

Let D be the single axiom of $I\Sigma_1$. Let

$$A' := (D \wedge A),\ B' := (D \wedge B),\ C' := (D \to C).$$

Then:

$$\mathsf{PRA} \nvdash A' \rhd_{\mathsf{PRA}} B' \to (A' \wedge \Box_{\mathsf{PRA}} C') \rhd_{\mathsf{PRA}} (B' \wedge \Box_{\mathsf{PRA}} C')$$

Hence M is not part of IL(PRA).

An even better example, is the following principle that is (i) in the intersection of ILM and ILP, but (ii) is not arithmetically valid in PRA.

$\mathsf{P}\Pi \vdash \phi \rhd \Diamond \psi \to \Box(\phi \rhd \Diamond \psi)$

To show that this principle is not PRA-valid, we need the following theorem due to Shavrukov (see [41]).

Theorem 9.1 *Suppose a consistent, finitely axiomatized, sequential theory F interprets a reflexive theory R that extends $I\Delta_0 + \mathsf{EXP}$. Then $\{P{\in}\Pi_1 \mid F \rhd (R+P)\}$ is complete Σ^0_3.*

We reason as follows. Clearly $I\Sigma^0_2 \rhd (\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA}))$. So $I\Sigma^0_2$ and $\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA})$ satisfy the conditions of Shavrukov's theorem. Hence,

$$\{P{\in}\Pi_1 \mid I\Sigma^0_2 \rhd (\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA}) + P)\}$$

is complete Σ^0_3. By a well-known theorem due independently to Goldfarb, Friedman and Harrington, there is a primitive recursive function F, transforming P to P', with $\mathsf{PRA} \vdash (\mathsf{Con}(\mathsf{PRA}){\wedge}P) \leftrightarrow \mathsf{Con}(\mathsf{PRA} + P')$. Hence

$$I\Sigma^0_2 \rhd (\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA}) + P) \Leftrightarrow I\Sigma^0_2 \rhd (\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA} + P'))$$

This tells us that $\{A \mid I\Sigma^0_2 \rhd (\mathsf{PRA} + \mathsf{Con}(\mathsf{PRA} + A))\}$ is complete Σ^0_3. Let D be the single axiom of $I\Sigma^0_2$. We can reformulate our insight as:

$$X := \{A \mid D \rhd_{\mathsf{PRA}} \Diamond_{\mathsf{PRA}} A)\} \text{ is complete } \Sigma^0_3.$$

On the other hand, if $\mathsf{P}\Pi$ would hold, by the soundness of PRA, X would be Σ^0_1. Quod non.

Remember that PRA is EA ($I\Delta_0 + \mathsf{EXP}$) plus the Π^0_2-induction rule. So it is reasonable to expect some analogies between PRA and the theories $\mathsf{EA} + \Pi^0_n\text{-}IR$. We have that $\mathsf{EA} + \Pi^0_n\text{-}IR$ is Σ_n-reflexive and hence we obtain analogues of the rule B, where we need suitable classes S_3 for $\mathsf{EA} + \Pi^0_3\text{-}IR$ and S_4 for $\mathsf{EA} + \Pi^0_n\text{-}IR$ with $n \geq 4$. Note that S_4 is simply all modal formulas.

Let $n \geq 3$. Clearly, $T := \mathsf{EA} + \Pi^0_n\text{-}IR$ extends $I\Sigma_1$. For $A, B \in \Sigma_2$, we have: $T \vdash A \rhd_T B \leftrightarrow A \rhd_{\Pi_1\text{-}\mathsf{cons}} B$. The interpretations of modal formulas used in the proof of the Hájek-Montagna arithmetical completeness theorem, in which it is shown that ILM is the the logic of Π_1-conservativity for Σ_1-sound extensions of $I\Sigma_1$, are all Δ_2. Hence *on these specific interpretations* Π_1-conservativity and interpretability coincide. It follows that every counterexample for ILM is a counterexample for $\mathsf{IL}(T)$. In other words: $\mathsf{IL}(T) \subseteq \mathsf{ILM}$. It is open whether $\mathsf{IL}(\mathsf{PRA})$ is a sublogic of ILM.

10 Other interpretations of $\rhd$

10.1 Partial Conservativity

Hájek and Montagna show that the logic of Π_1-conservativity of all extensions of $I\Sigma_1$ that do not prove their own iterated inconsistencies is ILM. See [23]. Japaridze gives a proof that does not use simplified models. See [19]. Konstantin Ignatiev characterizes the logics for partial conservativity for the classes Π_n for $n \geq 2$ and Σ_n for $n \geq 3$. He characterizes the closed fragment of the logic for Σ_1-conservativity. In these logics the principle $\mathsf{J5}$ is conspicuously absent. This suggests a more comprehensive study starting with IL_0 (= IL minus $\mathsf{J5}$).

10.2 Efficient Interpretability

It is typical of the interpretations in the literature that the proofs of the interpretations of the axioms of the target theory are fairly simple. So all these interpretations are in some sense 'efficient'. Let's explicate efficiency as the demand that there is a polynomial P such that for every axiom A there is a proof p of the interpretation of A where the size (= number of symbols) of p is bounded by P of the size of A. In other words the sizes of the proofs are bounded by an $\mathsf{P\text{-}Time}$ computable function. Efficient interpretability is this sense is studied by Rineke Verbrugge in [49],[50]. In [49] it is shown that ILM is sound and complete for arithmetical interpretations in PA interpreting $\rhd$ as feasible interpretability. In [50] it is shown that feasible interpretability *over* PA is Σ_2-complete. For more information on complexity, see subsection A.2.

10.3 Functional Relative Consistency

Christian Bennet studied in his thesis [4] the following notion of *strong relative consistency.* Let T and U be extensions of PA in the language of PA. Let PR be the set of primitive recursive terms. Define:

$$T \rhd_{\mathsf{src}} U :\Leftrightarrow \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PA}}(\forall x\,(\mathsf{Proof}_U(x,\bot) \to \mathsf{Proof}_T(tx,\bot))).$$

It was pointed out in [55] that the logic of $A \rhd_{\mathsf{rrc,PA}} B$ is ILP. The proof rests on the following 'Friedman Characterization' of $\rhd_{\mathsf{src}}$.

$$\mathsf{PA} \vdash T \rhd_{\mathsf{src}} U \leftrightarrow \Box_{\mathsf{PRA}}(\Diamond_T \top \to \Diamond_U \top).$$

The proof follows a suggestion of Kreisel. It uses the PA-verifiable fact that the provably recursive functions of PRA are precisely the primitive recursive ones. We have, in PA, writing $(*)$ for $\Box_{\mathsf{PRA}}(\Diamond_T \top \to \Diamond_U \top)$:

$$\begin{array}{rll}
(*) & \to & \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PRA}}\forall x\ (\mathsf{Proof}_U(x,\bot) \to \mathsf{Proof}_T(tx,\bot)) \\
& \to & \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PA}}\forall x\ (\mathsf{Proof}_U(x,\bot) \to \mathsf{Proof}_T(tx,\bot)) \\
& \to & \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PRA}}\Box_{\mathsf{PA}}\forall x\ (\mathsf{Proof}_U(x,\bot) \to \mathsf{Proof}_T(tx,\bot)) \\
& \to & \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PRA}}\forall x\ (\Box_{\mathsf{PA}}\mathsf{Proof}_U(x,\bot) \to \Box_{\mathsf{PA}}\mathsf{Proof}_T(tx,\bot)) \\
& \to & \exists t{\in}\mathsf{PR}\ \Box_{\mathsf{PRA}}\forall x\ (\mathsf{Proof}_U(x,\bot) \to (\neg\mathsf{Proof}_T(tx,\bot) \to \Box_{\mathsf{PA}}\bot)) \\
& \to & (*)
\end{array}$$

Clearly there are many variations on the theme of strong relative consistency. Vary the theories involved. Use $\mathsf{P\text{-}Time}$ or Kalmar Elementary instead of Primitive Recursive. Etcetera.

Note that the Friedman characterization for interpretability in finitely axiomatized sequential theories says that $T \rhd U$ iff U is strongly cutfree (tableaux) consistent relative to T where we use Kalmar Elmentary instead of Primitive Recursive.

10.4 Constructive Logic

It is unknown what the provability logic $\mathsf{L(HA)}$ of HA (Heyting Arithmetic) is. It is even compatible with our present state of knowledge that it is complete Π^0_2. $\mathsf{L(HA)}$ is a modal logic that contains the intuitionistic propositional calculus, IPC, plus the principles $\mathsf{L1\text{-}4}$ of Löb's Logic. In addition it contains many others like:

- $\Box\neg\neg\Box\phi \to \Box\Box\phi$
- $\Box(\neg\neg\Box\phi \to \Box\phi) \to \Box\Box\phi$
- $\Box(\phi \vee \psi) \to \Box(\phi \vee \Box\psi)$ (Leivant's Priciple)

The notion of Σ_1-*preservativity* turns out to be very useful in the study of provability principles in HA. Remember that Σ_1-preservativity is given by:

$$A \rhd_{\Sigma\text{-pres},\mathsf{HA}} B :\Leftrightarrow \forall S{\in}\Sigma_1\text{-sentences } (\Box_{\mathsf{HA}}(S \to A) \to \Box_{\mathsf{HA}}(S \to B)).$$

It is well possible that it is easier to formulate the logic of Σ_1-preservativity, then the logic of provability alone. The following principles are an attractive subsystem of Σ_1-preservativity logic of HA. Of course the basic logic here is IPC.

$\iota 1$ $\vdash \phi \to \psi \ \Rightarrow \vdash \phi \rhd \psi$

$\iota 2$ $\vdash (\phi \rhd \psi \wedge \psi \rhd \chi) \to \phi \rhd \chi$

$\iota 3$ $\vdash (\chi \rhd \phi \wedge \chi \rhd \psi) \to \chi \rhd (\phi{\wedge}\psi)$

$\iota 4$ $\vdash (\Box\phi \to \phi) \rhd \phi$

$\iota 5$ $\vdash (\phi \rhd \chi \wedge \psi \rhd \chi) \to (\phi{\vee}\psi) \rhd \chi$

The surprizing new principle is $\iota 5$. This principle corresponds via contraposition to the IL-invalid principle: $\vdash (\chi \rhd \phi \wedge \chi \rhd \psi) \to \chi \rhd (\phi{\wedge}\psi)$. One easy counterexample to this last principle was provided by the existence of oreysentences. It is easy to see that in the intuitionistic case we have oreysentences O for $\rhd_{\Sigma\text{-pres},\mathsf{HA}}$ too. These have the property: $\mathsf{O} \rhd \bot$ and $\neg\mathsf{O} \rhd \bot$. Applying $\iota 5$ gives us: $(\mathsf{O} \vee \neg\mathsf{O}) \rhd \bot$. This is not a contradiction, but just a testimony of the non-derivability of excluded third in HA!

Regrettably the beautiful logic $\iota 1\text{-}5$ does not exhaust the Σ_1-preservativity principles valid in HA. The reader is referred to [52],[59],[60] for more information.

11 Extensions of the language

11.1 Witness comparisons

Hájek and Montagna characterize the Π_1-conservativity logic for extensions of $I\Sigma_1$ with the witness comparison order added. See their [24]. This is an important result since a lot of interesting arguments can be formalized in this logic. De Jongh and Pianigiani apply the result of Hájek and Montagna to solve a problem posed by Guaspari. See [11].

Švejdar, in his fundamental paper [45], considers variations on the provability predicate that give a different meaning to the witness comparison formulas. It is open to give arithmetical completeness theorems for the interpretations he considers.

11.2 Propositional constants

The simplest kind of extension of the languages of provability and interpretability logic is, of course, extension with one propositional constant representing some significant statement in the intended target theory. For example one could extend the interpretability logic $\mathsf{IL}(\mathsf{ZF})$ with a constant for the Continuum Hypothesis CH. Let's call the logic obtained by extending the logic of a target theory T by adding a constant for a designated sentence A: $\mathsf{IL}(T, A)$.

The only example of a result along this line that I know of is the characterization of the *closed fragment* of $\mathsf{IL}(I\Delta_0+\Omega_1,\mathsf{EXP})$ in [57]. In this fragment many salient facts about the relationship between the *weak* theory and the *strong* axiom can be formulated, like Paris and Wilkie's result that $\mathsf{EXP} \rhd \Diamond\top$, Solovay's result that $\top \rhd \neg\mathsf{EXP}$ and my own result that $\Diamond\top \rhd \mathsf{EXP}$.

11.3 Weak interpretability and tolerance

Giorgi Japaridze studies the notions of weak interpretability and tolerance. Weak interpretability of U in T means that U is interpretable in some consistent extension of T. Tolerance is a generalization of this notion to the n-ary case. The reader is referred to Japaridze's papers [17] and [18].

11.4 Σ_1-Interpolability

Konstantin Ignatiev, in his paper [28], characterizes the logic of Σ_1-interpolability over PA. Σ_1-interpolability over PA is defined as follows.

- $A \twoheadrightarrow_{\mathsf{PA}} B :\leftrightarrow \exists S{\in}\Sigma^0_1\text{-sentences } (\Box_{\mathsf{PA}}(A \to S) \wedge \Box_{\mathsf{PA}}(S \to B))$.

Interpolability can be viewed as a kind of dual of Π_1-conservativity. In the language predicates like *being provably equivalent to a* Σ_1*-sentence* and *weak interpretability* can be expressed. Ignatiev shows that the logic of Σ_1-interpolability satisfies interpolation —in contrast to ILM.

The problem of characterizing the combined logic of interpretability and interpolability is open. We do not know whether this combined logic satisfies interpolation.

11.5 Feferman's Predicate

The Feferman Predicate for a reflexive theory T is the predicate $\triangle$ that is defined as follows.

- $\triangle_T A :\leftrightarrow \exists x\,(\Box_{T,x} A \wedge \Diamond_{T,x} \top)$.

Here, for theories that contain SUPEXP, $\Box_{T,x}A$ means provability of A from the axioms of T with gödelnumbers smaller than x. (If the theory does not contain SUPEXP one must also restrict the formulas used in the proof to those with complexity $\leq |x|$. See also appendix C.) The Feferman Predicate is a provability predicate with 'built in consistency'. It is an indispensable technical tool for constructing interpretations via the Henkin construction. The gödelsentence Feferman provability in PA is an oreysentence for PA.

The first systematic study of the Feferman Predicate for PA in provability logic was made by the author in his paper [54]. However, Volodya Shavrukov was the first to provide a modal system for combined ordinary provability and Feferman provability that is arithmetically complete for interpretations in PA. In the interpretation of Shavrukov's system one makes use of a *specific choice* of the enumeration of the axioms of PA. See [40].

Give the important connection between Feferman provability and interpretability, it would be interesting to see a characterization of the combined logic of these two notions.

References

[1] L.D. Beklemishev. Iterated local reflection vs iterated consistency. Technical Report LGPS 118, Department of Philosophy, Utrecht University, 1994.

[2] L.D. Beklemishev. Induction rules, reflection principles and provably recursive functions. Technical Report LGPS 168, Department of Philosophy, Utrecht University, 1996.

[3] L.D. Beklemishev. Parameter free induction and reflection. Technical Report LGPS 171, Department of Philosophy, Utrecht University, 1996.

[4] C. Bennet. *On some orderings of extensions of arithmetic.* Department of Philosophy, University of Göteborg, 1986.

[5] A. Berarducci. The interpretability logic of Peano arithmetic. *The Journal of Symbolic Logic*, 55:1059–1089, 1990.

[6] A. Berarducci and R. Verbrugge. On the provability logic of bounded arithmetic. *Annals of Pure and Applied Logic*, 61:75–93, 1993.

[7] G. Boolos. *The logic of provability.* Cambridge University Press, 1993.

[8] S. Buss. *Bounded Arithmetic.* Bibliopolis, Napoli, 1986.

[9] P. Clote and J. Krajíček, editors. *Arithmetic, Proof Theory and Computational Complexity.* Oxford University Press, Oxford, 1993.

[10] P. Clote and J. Remmel, editors. *Feasible Mathematics II.* Birkhaüser, Boston, 1994.

[11] D. de Jongh and D. Pianigiani. Solution of a problem of David Guaspari. *Studia Logica*, 10000:10000–100001, 1996.

[12] D. de Jongh and F. Veltman. Provability logics for relative interpretability. In *[36]*, pages 31–42, 1990.

[13] D. de Jongh and A. Visser. Explicit fixed points in interpretability logic. *Studia Logica*, 50:39–50, 1991.

[14] M. de Rijke. Bi-unary interpretability logic. Technical Report Report X-90-12, ITLI, University of Amsterdam, 1990.

[15] M. de Rijke. A note on the interpretability logic of finitely axiomatized theories. *Studia Logica*, 50:241–250, 1991.

[16] M. de Rijke. Unary interpretability logic. *The Notre Dame Journal of Formal Logic*, 33:249–272, 1992.

[17] G. Dzhaparidze (Japaridze). The logic of linear tolerance. *Studia Logica*, 51:249–277, 1992.

[18] G. Dzhaparidze (Japaridze). A generalized notion of weak interpretability and the corresponding logic. *Annals of Pure and Applied Logic*, 61:113–160, 1993.

[19] G. Dzhaparidze (Japaridze). A simple proof of arithmetical completeness for Π_1-conservativity logic. *The Notre Dame Journal of Formal Logic*, 35:346–354, 1994.

[20] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 49:35–92, 1960.

[21] M.J. Greenberg. *Euclidean and Non-Euclidean Geometries, 3d edition.* Freeman, 1996.

[22] D. Guaspari. Sentences implying their own provability. *The Journal of Symbolic Logic*, 48:777–789, 1983.

[23] P. Hájek and F. Montagna. The logic of Π_1-conservativity. *Archiv für Mathematische Logik und Grundlagenforschung*, 30:113–123, 1990.

[24] P. Hájek and F. Montagna. The logic of Π_1-conservativity continued. *Archiv für Mathematische Logik und Grundlagenforschung*, 32:57–63, 1992.

[25] P. Hájek and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1991.

[26] P. Hájek and V. Švejdar. A note on the normal form of closed formulas of interpretability logic. *Studia Logica*, 50:25–38, 1991.

[27] D. Hilbert and S. Cohn-Vossen. *Anschauliche Geometrie*. Wissenschaftliche Buchgesellschaft, Darmstadt, 1973.

[28] K.N. Ignatiev. The provability logic of Σ_1-interpolability. *Annals of Pure and Applied Logic*, 64:1–25, 1993.

[29] M.B. Kalsbeek. An orey sentence for predicative arithmetic. Technical Report Report X-89-01, ITLI, University of Amsterdam, 1989.

[30] M.B. Kalsbeek. Towards the interpretability logic of $i\delta_0 + \mathsf{exp}$. Technical Report LGPS 61, Department of Philosophy, Utrecht University, 1991.

[31] P. Lindström. On partially conservative sentences and interpretability. *Proceedings of the American Mathematical Society*, 91:436–443, 1984.

[32] L. Manevitz and J. Stavi. Δ^0_2 operators and alternating sentences in arithmetic. *The Journal of Symbolic Logic*, 45:144–154, 1980.

[33] J. Mycielski, P. Pudlák, and A.S. Stern. *A lattice of chapters of mathematics (interpretations between theorems)*, volume 426 of *Memoirs of the American Mathematical Society*. AMS, Providence, Rhode Island, 1990.

[34] E. Nelson. *Predicative arithmetic*. Princeton University Press, Princeton, 1986.

[35] J.B. Paris and A. Wilkie. On the scheme of of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, 35:261–302, 1987.

[36] P.P. Petkov, editor. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria*. Plenum Press, Boston, 1990.

[37] P. Pudlák. Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, 50:423–441, 1985.

[38] P. Pudlák. On the length of proofs of finitistic consistency statements in finitistic theories. In J.B. et al Paris, editor, *Logic Colloquium '84*, pages 165–196. North–Holland, Amsterdam, 1986.

[39] V.Yu. Shavrukov. The logic of relative interpretability over Peano arithmetic (in Russian). Technical Report Report No.5, Stekhlov Mathematical Institute, Moscow, 1988.

[40] V.Yu. Shavrukov. A smart child of Peano's. *The Notre Dame Journal of Formal Logic*, 35:161–185, 1994.

[41] V.Yu. Shavrukov. Interpreting reflexive theories in finitely many axioms. Technical Report LGPS 138, Department of Philosophy, Utrecht University, 1995.

[42] C. Smoryński. *Self-reference and modal logic.* Springer-Verlag, 1985.

[43] R.M. Solovay. On interpretability in set theories. Unpublished manuscript, 1976.

[44] C. Strannegård. Interpretability over finitely axiomatized theories. Unpublished manuscript, 1996.

[45] V. Švejdar. Modal analysis of generalized rosser sentences. *The Journal of Symbolic Logic*, 48:986–999, 1983.

[46] V. Švejdar. Some independence results in interpretability logic. *Studia Logica*, 50:29–38, 1991.

[47] A. Tarski, A. Mostowski, and R.M. Robinson. *Undecidable theories.* North–Holland, Amsterdam, 1953.

[48] L.C. Verbrugge. *Efficient metamathematics.* ILLC-disseration series 1993-3, Amsterdam, 1993.

[49] L.C. Verbrugge. Feasible interpretability. In *[9]*, pages 387–428. 1993.

[50] L.C. Verbrugge. The complexity of feasible interpretability. In *[10]*, pages 429–447. 1994.

[51] A. Visser. The provability logics of recursively enumerable theories extending Peano arithmetic at arbitrary theories extending Peano arithmetic. *Journal of Philosophical Logic*, 13:97–113, 1984.

[52] A. Visser. Evaluation, provably deductive equivalence in heyting's arithmetic of substitution instances of propositional formulas. Technical Report LGPS 4, Department of Philosophy, Utrecht University, 1985.

[53] A. Visser. Preliminary notes on interpretability logic. Technical Report LGPS 29, Department of Philosophy, Utrecht University, 1988.

[54] A. Visser. Peano's smart children: A provability logical study of systems with built-in consistency. *Notre Dame Journal of Formal Logic*, 30:161–196, 1989.

[55] A. Visser. Interpretability logic. In *[36]*, pages 175–209, 1990.

[56] A. Visser. The formalization of interpretability. *Studia Logica*, 51:81–105, 1991.

[57] A. Visser. An inside view of exp. *The Journal of Symbolic Logic*, 57:131–165, 1992.

[58] A. Visser. The unprovability of small inconsistency. *Archive for Mathematical Logic*, 32:275–298, 1993.

[59] A. Visser. Propositional combinations of Σ–sentences in heyting's arithmetic. Technical Report LGPS 117, Department of Philosophy, Utrecht University, 1994.

[60] A. Visser, J. van Benthem, D. de Jongh, and G. Renardel. *NILL*, a study in intuitionistic propositional logic. In A. et al. Ponse, editor, *Modal Logic and Process Algebra, a Bisimulation Perspective*, pages 289–326. CSLI Publications, Lecture Notes, no. 53, 1995.

[61] M. Vuković. Some correspondences of priinciples of interpretability logic. *Glasnik Matematički*, 31(51):193–200, 1996.

[62] D. Zambella. On the proofs of arithmetical completeness of interpretability logic. *The Notre Dame Journal of Formal Logic*, 35:542–551, 1992.

A Some pointers to further work

A.1 Restrictions of the language

Maarten de Rijke studies unary interpretability logic, i.e. the logic of the predicate $(\top \rhd \phi)$. It turns out that his *il*, *ilp* and *ilm* all satisfy interpolation. The reader is referred to his papers [14], [16].

A.2 Complexity

Robert Solovay (in his unpublished [43]) and Per Lindström (in [31]) prove independently that $\{S{\in}\Sigma_1 \mid \top \rhd_T S\}$, where T is a Σ_1-sound reflexive theory, is Π_2-complete. Volodya Shavrukov shows, in his [41], that $\{P{\in}\Pi_1 \mid \mathsf{GB} \rhd (\mathsf{ZF} + P)\}$ and $\{S{\in}\Sigma_1 \mid \mathsf{GB} \rhd (\mathsf{ZF} + S)\}$ are Σ_3-complete. This shows that things are as bad as they can get.

A.3 Embeddings of Algebras

Claes Strannegård, in his unpublished [44] generalizes the completeness theorem for ILP to a result on embeddings of algebras analogous to Shavrukov's result on embedding diagonalizable algebras.

B A list of problems in Interpretability Logic

For completeness sake I added, what I take to be, the two major open problems of *Provability Logic* as problems 1 and 2 to my list. The major open problems of Interpretability Logic are 5,7,8. Problems 6,12,14 are purely modal.

1. What is the provability logic of $I\Delta_0 + \Omega_1$? What is the provability logic of Buss' S^1_2? Remarks:

   - At present it cannot be excluded that the problem is connected to problems in complexity theory.
   - Even if the problem is mainly arithmetical, it also has modal aspects. E.g., formulate a plausible extension of Löb's Logic of which it is not immediately evident that it cannot be the provability logic of $I\Delta_0 + \Omega_1$.
   - It would even be possible that the logics of $I\Delta_0 + \Omega_1$ and S^1_2 differ. That would be rather surprising. Can it be proved that these logics are the same?

   Clearly, at this stage it is not wise to ask about the interpretability logic of $I\Delta_0 + \Omega_1/\mathsf{S}^1_2$. On the other hand, looking at the more difficult problem could help to suggest provability principles extending Löb's Logic. Fro some information, see subsection 9.1.

2. What is the provability logic of Heyting's Arithmetic HA and of related theories (like HA+ECT_0, HA+M_{PR})? See subsection 10.4.

3. Give a characterization of interpretability that works for all reasonable arithmetical theories. See appendix C.

4. Give a construction of an oreysentence that works for all reasonable arithmetical and sequential theories. See subsection 3.5. An easier variant is the same question for *local* interpretability.

5. What is the interpretability logic of all reasonable arithmetical theories. I conjecture that this logic is ILW^*. See section 6. See also [56]. Note that there are all kinds of variants. What is the interpretability logic of all finitely axiomatized extensions of PRA? What is the interpretability logic of $I\Sigma_1$ for all possible choices of the set of designated numbers? Etcetera.

6. Axiomatize $\mathsf{IL}(I\Delta_0 + \mathsf{EXP})$. See subsection 9.2.

7. What is $\mathsf{IL}(\mathsf{PRA})$? What is $\mathsf{IL}(I\Sigma_n\text{-}IR)$ for $n \geq 2$? Here $\mathsf{IL}(I\Sigma_n\text{-}IR)$ is $I\Delta_0 + \mathsf{EXP}$ plus the Σ_n-induction rule. See subsection 9.3.

8. What is the logic for Π_1-conservativity in PRA? See subsections 9.3 and 10.1.

9. What are the logics of Σ_1- and of Σ_2-conservativity over PA? See subsection 10.1.

10. What are the interpretability logics for the language with witness comparisons, where the 'proof-predicate' is interpreted using Švejdar's interpretations II and III? See [45], p989,990. For a treatment of Švejdar's interpretation I, see [24]. As far as I know these problems are open even for the language with $\Box$, but without $\rhd$. See also subsection 11.1.

11. What is the logic of interpretability, extended with the Feferman predicate in the style of Shavrukov. See subsection 11.5.

12. Give a modal completeness theorem for ILW^*. See subsection 6.

13. Do IL, ILW, ILP satisfy interpolation?

14. Does the logic for interpretability and Σ_1-interpolability of Ignatiev satisfy interpolation? See subsection 11.4.

15. What is the logic of the predicate $(\mathsf{GB} + A) \rhd (\mathsf{ZF} + B)$, where A, B are in the language of ZF and where our basis theory is ZF? (Note that $(\mathsf{ZF} + A) \rhd (\mathsf{GB} + B)$ is equivalent to $\Box_{\mathsf{ZF}}(A \to \Diamond_{\mathsf{GB}} B)$.)

16. Is it possible to extend the approach of appendix D to a different basic theory than PA. One problem is that the convenient fact that internally definable models are endextensions disappears.

C The Orey-Hájek characterization and other matters

In this appendix we prove a refined version of theorem 3.2. We will only consider sequential theories. In this section, restricted provability, Prov_n and $\Box_n$, will mean that we restrict the axioms used to those with gödelnumber $\leq n$ and the proofs to those only involving formulas A of complexity $\rho(A) \leq |n|$. Here $|n| := \mathsf{entier}({}^2\mathsf{log}(n+1))$. Our notion of complexity ρ counts depth of quantifier changes. We will make the natural assumption that $\rho(A) \leq |\#A|$. We will say that a theory V is *r-reflexive* if it proves, for all n, $\mathsf{Con}_n(V)$. In the presence of the axiom SUPEXP in T and U we can drop the restriction on the complexity of the formulas involved in te proofs, since we can prove cutelimination for predicate logic. In other words, for theories proving SUPEXP reflexivity and r-reflexivity conincide. We will formulate our result below partly in terms of $\forall\Pi_1^b$-conservativity, rather than Π_1^0-conservativity. In the presence of the axiom EXP in T and U $\forall\Pi_1^b$-conservativity coincides with ordinary Π_1^0-conservativity. We write $T \rhd_{\mathsf{OH}} U$ for: $\forall x\ \Box_T \Diamond_{U,x} \top$. Note that T is r-reflexive iff $T \rhd_{\mathsf{OH}} T$.

Lemma C.1 Let theories T, U, W be given.

1. Suppose that W proves Π^b_2-completeness[17] and that T is W-verifiably r-reflexive. Then, $W \vdash T \rhd_{\mathsf{loc}} U \to T \rhd_{\mathsf{OH}} U$.

2. $W \vdash T \rhd_{\mathsf{OH}} U \to T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$

3. Let U be W-verifiably r-reflexive. Then, $W \vdash T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U \to T \rhd_{\mathsf{OH}} U$.

4. $W \vdash T \rhd_{\mathsf{OH}} U \to T \rhd U$

5. $W \vdash T \rhd U \to T \rhd_{\mathsf{loc}} U$

❑

Proof

(1) Suppose that W proves Π^b_2-completeness and that T is W-verifiably r-reflexive. Reason in W. Suppose $T \rhd_{\mathsf{loc}} U$. Let $\alpha_{U,x}$ be the formula defining the set of axioms of U with gödelnumber $\leq x$. Since, we did not assume Σ^0_1-collection, we have to stipulate that '$T \rhd_{\mathsf{loc}} U$' means:

$$\forall x\, \exists \mathcal{K}, y\, \forall a{\in}\alpha_{U,x}\, \exists p{\leq}y\ \mathsf{Proof}_T(p, a^{\mathcal{K}}).$$

Fix x. By Π^b_2-completeness we have: $\Box_T(\forall a{\in}\alpha_{U,x}\, \exists p{\leq}y\ \mathsf{Proof}_T(p, a^{\mathcal{K}}))$. We can use this last fact, inside $\Box_T$, to transform an x-proof q in U of, say, b into a proof q^* of $b^{\mathcal{K}}$ in T. The T-axioms involved in q^* will be bounded by y. A subproof of q^* that verifies the interpretation of a U-axiom will only involve formulas with complexity bounded by $|y|$. Those parts of q^* that 'simulate' q will only contain T-formulas obtained by $\mathcal{K}$-translating U-formulas of complexity at most $|x|$. These T-formulas will have complexities bounded by $|x| + |\mathcal{K}|$. Thus we can find a z such that: $\Box_T(\Box_{U,x}\bot \to \Box_{T,z}\bot)$. Ergo, $\Box_T\Diamond_{U,x}\top$.

To prove (2), reason in W. Suppose $\forall x\ \Box_T\Diamond_{U,x}\top$ and $\Box_U P$, where $P \in \forall\Pi^b_1$. We can find an x such that $\Box_T\Box_{U,x}P$ and $\Box_T(\neg P \to \Box_{U,x}\neg P)$. (The last fact holds, since all theories containing S^1_2 prove $\exists\Sigma^b_1$-completeness, where the complexity of the witnessing proof is linear in the complexity of the $\exists\Sigma^b_1$-formula.) Combining we find: $\Box_T P$. (Note that, from a bound of on the U-proofs of P, we can extract a bound on the T-proofs of P. Thus proving 'smooth' $\forall\Pi^b_1$-conservativity.)

(3) Suppose that U is W-verifiably r-reflexive. Reason in W. Suppose $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$. Consider any x. Since $\Box_U\Diamond_{U,x}\top$, we have immediately, $\Box_T\Diamond_{U,x}\top$.

To prove (4), we use the Henkin construction as described in [56] on the Feferman predicate $\Box^*_U A :\leftrightarrow \exists x\,(\Box_{U,x}A \wedge \Diamond_{U,x}\top)$. This gives us an interpretation $\mathcal{K}$ of U in T.

Finally, (5) is trivial. ❑

[17] It is sufficient for Π^b_2-completeness that W proves EXP.

In the next theorem we just harvest the fruits of the preceding lemma.

Theorem C.2

- *Suppose that W proves Π^b_2-completeness and that T is W-verifiably r-reflexive. Then, $T \rhd_{\mathsf{loc}} U$, $T \rhd U$ and $T \rhd_{\mathsf{OH}} U$ are W-provably equivalent.*
- *Suppose that U is W-verifiably r-reflexive, then $T \rhd_{\mathsf{OH}} U$ and $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$ are W-provably equivalent.*
- *Suppose that W proves Π^b_2-completeness and that T, U are W-verifiably r-reflexive. Then $T \rhd_{\mathsf{loc}} U$, $T \rhd U$, $T \rhd_{\mathsf{OH}} U$ and $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$ are W-provably equivalent.*

Following an idea of Lev Beklemishev we show the 'necessity'[18] of the conditions on T, U of lemma C.1. We have been looking at four notions for comparing theories. These give rise to 16 implications of the form $T \rhd_{\mathsf{X}} U \to T \rhd_{\mathsf{Y}} U$. Since we want to show the necessity of the conditions, it is sufficient to consider *the weakest non-trivial implications*. These turn out to be $T \rhd_{\mathsf{loc}} U \to T \rhd U$, $T \rhd U \to T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$ and $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U \to T \rhd_{\mathsf{loc}} U$. We can replace our last implication by the even weaker non-trivial implication: $T \rhd_{\Pi^0_1\text{-}\mathsf{con}} U \to T \rhd_{\mathsf{loc}} U$. We will say that T is **r-reflexive* if T is r-reflexive for some choice of the designated natural numbers of T. One can show that if T is *r-reflexive and if $\mathcal{N}$ is any choice of the natural numbers for T, then there is a definable $\mathcal{N}$-cut $\mathcal{I}$, such that $\langle T, \mathcal{I}\rangle$ is r-reflexive.

Theorem C.3 *We work in a suffiently rich metatheory W.*

1. *Let T be given. Suppose, for all U, if $T \rhd_{\mathsf{loc}} U$, then $T \rhd U$. Then T is *r-reflexive.*
2. *Let T be given. Suppose, for all U, if $T \rhd U$, then $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$. Then T is r-reflexive.*
3. *Let U be given. Suppose, for all T, if $T \rhd_{\Pi^0_1\text{-}\mathsf{con}} U$, then $T \rhd_{\mathsf{loc}} U$. Suppose also that $\Box_U \mathsf{EXP}$. Then U is r-reflexive.*

Proof

(1) Suppose for all U, if $T \rhd_{\mathsf{loc}} U$, then $T \rhd U$. Take

$$U^* := I\Delta_0 + \Omega_1 + \{\Diamond_{T,n}\top \mid n \in \omega\}.$$

[18] As the reader will see the Beklemishev explication of 'necessity' focusses on the question of characterizing properties of the form $P(T) :\Leftrightarrow \forall U\ (T \rhd_1 U \Rightarrow T \rhd_2 U)$ or $Q(U) :\Leftrightarrow \forall T\ (T \rhd_1 U \Rightarrow T \rhd_2 U)$. Of course, this just one possible way of explicating the question of necessity. E.g. some further restriction of the quantifier in the definition of P could change the picture. Etcetera.

By familiar arguments —see e.g [58]— we have, $T \rhd_{\mathsf{loc}} U^*$. By assumption, for some $\mathcal{K}$, $\mathcal{K} : T \rhd U^*$. It is not difficult to see that we can always replace $\mathcal{K}$ by a T-cut. Take the new numbers as given by $\mathcal{K}$. So T will be W-verifiably r-reflexive w.r.t $\mathcal{K}$.

(2) Suppose that for all U, if $T \rhd U$, then $T \rhd_{\forall\Pi^b_1\text{-}\mathsf{con}} U$. We have, for any n, by familiar arguments, that $T \rhd (I\Delta_0 + \Omega_1 + \Diamond_{T,n}\top)$. Hence, by conservativity, $\Box_T \Diamond_{T,n}\top$. Ergo T is r-reflexive.

(3) Suppose that, for all T, if $T \rhd_{\Pi^0_1\text{-}\mathsf{con}} U$ then $T \rhd_{\mathsf{loc}} U$. Suppose also that U contains EXP. Let T^* be axiomatized by $I\Delta_0 + \Omega_1$, plus the Π^0_1-consequences of U. Using a version of Craig's trick we can make our axiomset Σ^b_1-defined. Clearly, $T^* \rhd_{\Pi^0_1\text{-}\mathsf{con}} U$, so $T^* \rhd_{\mathsf{loc}} U$. Fix n. By familiar arguments, for some U-cut $\mathcal{I}$, $\Box_U \Diamond^{\mathcal{I}}_{U,n}\top$. Let U_0 be a finite subtheory of U such that $\Box_{U_0}$ "$\mathcal{I}$ is a cut" and $\Box_{U_0} \Diamond^{\mathcal{I}}_{U,n}\top$. Ex hypothesi, for some $\mathcal{K}$, $\mathcal{K} : T^* \rhd U_0$. We can find a T^*-cut $\mathcal{J}$ such that on $\mathcal{J}$ there is a T^*-definiable isomorphism between $\mathcal{J}$ and a T^*-definable 'external cut' of $\mathcal{K}$. We can choose $\mathcal{J}$ so small that it image on the $\mathcal{K}$-side is in $\mathcal{I}$. We may conclude, by the downward persistence of Π^0_1-sentences, $\Box_{T^*} \Diamond^{\mathcal{J}}_{U,n}\top$. T^* is Π^0_1-axiomatized over $I\Delta_0 + \Omega_1$, so by the work of Paris and Wilkie (see [35], see also lemma 4.1 of [57], for a sharp version) it follows that: $\Box_{T^*+\mathsf{EXP}} \Diamond_{U,n}\top$. Hence, $\Box_U \Diamond_{U,n}\top$. (For the application of lemma 4.1, we need that our metatheory contains SUPEXP.) ❑

Note that we cannot get around asking something like "U proves EXP" in (3), since without this assumption $I\Delta_0 + \Omega_1$ also satisfies our condition on U. But $I\Delta_0 + \Omega_1$ is not r-reflexive.

We end this section with some remarks on local interpretability. Let's again work in a convenient meta theory, like the de luxe $I\Delta_0 + B\Sigma_1 + \mathsf{SUPEXP}$. Let $\mathsf{RCON}(T) := \{\Diamond_{T,n}\top \mid n \in \omega\}$ and $T^{\mathsf{rc}} := I\Delta_0 + \Omega_1 + \mathsf{RCON}(T)$ and $T^{\mathsf{rce}} := I\Delta_0 + \mathsf{EXP} + \mathsf{RCON}(T)$. By the considerations of e.g. [58], we have: $T \equiv_{\mathsf{loc}} T^{\mathsf{rc}}$. So we find, applying some ideas of Paris and Wilkie (see [35] and lemma 4.1 of [57]):

$$\begin{array}{rcl} T \rhd_{\mathsf{loc}} U & \Leftrightarrow & T^{\mathsf{rc}} \rhd_{\mathsf{loc}} U^{\mathsf{rc}} \\ & \Leftrightarrow & \forall n\, \exists \mathcal{I} {\in} (I\Delta_0 + \Omega_1)\text{-cuts}\ \Box_{T^{\mathsf{rc}}} \Diamond^{\mathcal{I}}_{U,n}\top \\ & \Leftrightarrow & \forall n\ \Box_{T^{\mathsf{rce}}} \Diamond_{U,n}\top \end{array}$$

So, we may conclude: $T \rhd_{\mathsf{loc}} U \Leftrightarrow \forall n\ \Box_{T^{\mathsf{rce}}} \Diamond_{U,n}\top$. In a different formulation:

$$T \rhd_{\mathsf{loc}} U \Leftrightarrow \forall n\, \exists k\ \Box_{I\Delta_0+\mathsf{EXP}}(\Diamond_{T,k}\top \to \Diamond_{U,n}\top).$$

It is easy to see that T^{rc} is r-reflexive. So we have:

$$T \rhd_{\mathsf{loc}} U \Leftrightarrow T^{\mathsf{rc}} \rhd U \Leftrightarrow T^{\mathsf{rc}} \rhd U^{\mathsf{rc}}.$$

Let's write $[\![T]\!]$, $[\![T]\!]_{\mathsf{loc}}$ for, respectively, the interpretability type and the local interpretability type of T. We may consider the (local) interpretability types equipped with the obvious partial orders as degrees of (local) interpretability. Consider the mappings $\mathcal{F} : [\![T]\!]_{\mathsf{loc}} \mapsto [\![T^{\mathsf{rc}}]\!]$ and $\mathcal{G} : [\![T]\!] \mapsto [\![T]\!]_{\mathsf{loc}}$. Evidently the pair $\langle \mathcal{F}, \mathcal{G} \rangle$ is an adjunction between the degrees of local interpretability and the degrees of interpretability. We also see that the degrees of local interpretability of arbitrary theories are isomorphic to the degrees of interpretability of theories that are *r-reflexive. Note that these results hold for sequential theories that interpret, say, $I\Delta_0 + \Omega_1$. Since even Robinson's Q interprets $I\Delta_0 + \Omega_1$, this is, I guess, a reasonably wide class.[19]

D A Kripke model for Arithmetic

It would be nice to generalize insights from provability logic and interpretability logic to theories without substantial coding machinery. One way to this could be to replace the coded modal operators by modal operators defined in terms of natural relations between models like *extension*. In this appendix I study some relations between models of PA to which we can make the coded operators correspond. Moreover we will briefly look at a non-codable operator. The results of this section are not in the literature. However, specialists in the field have been aware that some such elaboration is possible.

A internal model of a model $\mathcal{M}$ is a model that is definable in $\mathcal{M}$ by a relative interpretation $\mathcal{N}$. A restricted internal model is an internal model for which we have a truthpredicate N in $\mathcal{M}$, i.e., for any A with (appropriately coded) parameters in the internal model, $\mathcal{M} \models N(\underline{\#A}) \leftrightarrow A^{\mathcal{N}}$. Define:

1. $\mathcal{M} \preceq_1 \mathcal{N} :\Leftrightarrow \mathcal{N}$ is an extension of $\mathcal{M}$,
2. $\mathcal{M} \preceq_2 \mathcal{N} :\Leftrightarrow \mathcal{N}$ is an endextension of $\mathcal{M}$,
3. $\mathcal{M} \preceq_3 \mathcal{N} :\Leftrightarrow \mathcal{N}$ is an extension of $\mathcal{M}$
   isomorphic to an internal model of $\mathcal{M}$,
4. $\mathcal{M} \preceq_4 \mathcal{N} :\Leftrightarrow \mathcal{N}$ is an extension of $\mathcal{M}$
   isomorphic to a restricted internal model of $\mathcal{M}$.

We are going to treat the relations $\preceq_i$ as partial orders between models. This is somewhat awkward since the correct way to treat these matters is to work with a category with as morphisms the embeddings of the domain of $\mathcal{M}$ into the domain of $\mathcal{N}$. Such embeddings are not fixed with the models since —as is well known— many models have several different isomorphic submodels. There is no problem at all with treating these matters categorically.

[19] A defect of the degrees approach is, perhaps, that we abstract away in a rather radical manner from the information contained in individual interpretations and 'local interpretations' (defined in some appropriate way). So here is an open question: can we work in more informative categories and still preserve the significant result above providing an adjunction. It is easy to see that the most obvious approach does not work since there are many alternatives for our specific choice of defining T^{rc}, which are not isomorphic in the category of theories and interpretations.

However, for the purposes of this paper it is somewhat laborious. Here we will refrain from identifying isomorphic models but restrict ourselves to embeddings in the strict sense, i.e. embeddings that send an object to itself. Thus we obtain the desired effect of working with a partial ordering. However, we are forced to the tortuous definition of e.g. $\preceq_3$. Having noted this, we will henceforth ignore these subtleties and pretend, e.g., that $\preceq_3$ simply means *is an internal model of*, etcetera.[20]

We write $\mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}$ for the language of PA with parameters in $\mathcal{M}$.

Theorem D.1 *We have:*

- *For* $1 \leq i < j \leq 4$*:* $\prec_j \subseteq \prec_i$.
- *For* $A \in \mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}$*, we have* $\exists \mathcal{M} \prec_i \mathcal{N} \;\; \mathcal{N} \models A \Leftrightarrow \exists \mathcal{M} \prec_j \mathcal{N} \;\; \mathcal{N} \models A.$.

Proof

The first part is easy. For the second part is clearly suffices to show that

$$\exists \mathcal{M} \prec_1 \mathcal{N} \;\; \mathcal{N} \models A \Rightarrow \exists \mathcal{M} \prec_4 \mathcal{N} \;\; \mathcal{N} \models A.$$

Consider any models $\mathcal{M}$, $\mathcal{N}$ and suppose $\mathcal{M} \prec_1 \mathcal{N}$ and $\mathcal{N} \models A$. For each standard number n we have $\mathcal{N} \models \Diamond_{\mathsf{PA},n} A$. By Matiyasevič's Theorem $\Diamond_{\mathsf{PA},n} A$ is PA-equivalent with a purely universal sentence. Hence, for each standard n, $\mathcal{M} \models \Diamond_{\mathsf{PA},n} A$. Use the Henkin construction on the Feferman-consistency statement of $\mathsf{PA} + A$ to produce the desired restricted internal model. ❑

It is easy to see that the first two of the above inclusions are strict. Is there an example to show that the third inclusion is strict (even modulo isomorphism)? (If we drop the requirement of reflexivity, the identity interpretation would be a trivial example ... by Tarski's theorem on the undefinability of truth.) Define:

- $\mathsf{REF}_{\mathcal{M}} := \{(\Box_m C \to C) \mid C \in \mathcal{L}^{\mathcal{M}}_{\mathsf{PA}} \text{ and } m \in \mathcal{M}\}$,
- $\mathcal{M} \prec_{i+} \mathcal{N} :\Leftrightarrow \mathcal{M} \preceq_i \mathcal{N} \models \mathsf{REF}_{\mathcal{M}}$.
- $\mathcal{M} \preceq_{i,\mathcal{K}} \mathcal{N} :\Leftrightarrow \mathcal{K} \prec_{i+} \mathcal{M}$, $\mathcal{K} \prec_{i+} \mathcal{N}$ and $\mathcal{M} \preceq_i \mathcal{N}$

I feel that it is a defect that we have to use syntax to define $\prec_{i+}$. Is there a more structural characterization?

Theorem D.2 *We have:*

1. $\preceq_i$ *is a partial ordering.*

[20] Another way to circumvent working in a category is by unraveling. Instead of models we work with sequences $\mathbb{N} =: \mathcal{M}_0, f_1, \mathcal{M}_1, \ldots, f_n \mathcal{M}_n$, where f_{j+1} is an i-embedding of $\mathcal{M}_j$ into $\mathcal{M}_{j+1}$. Our accessibility relation is extension of sequences.

2. $\prec_{i+}$ *is transitive and antisymmetric.*

3. $\prec_{i+} \subseteq \preceq_i$.

4. $\mathcal{M} \preceq_i \mathcal{N} \prec_{i+} \mathcal{P} \to \mathcal{M} \prec_{i+} \mathcal{P}$.

5. $\mathcal{M} \prec_{i+} \mathcal{N} \Leftrightarrow \forall C{\in}\mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}\ (\mathcal{M} \models \Box_{\mathsf{PA}} C \Rightarrow \mathcal{N} \models C)$.

6. $\mathcal{M} \prec_{i+} \mathcal{N} \Rightarrow \forall C{\in}\mathcal{L}^{\mathcal{N}}_{\mathsf{PA}} \forall k{\in}\mathcal{K}\ (\mathcal{N} \models \Box_k C \to C)$.

Proof

We only treat 5. Suppose $\mathcal{M} \prec_{i+} \mathcal{N}$, $C{\in}\mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}$ and $\mathcal{M} \models \Box_{\mathsf{PA}} C$. It follows that, for some m in $\mathcal{M}$, $\mathcal{M} \models \Box_{\mathsf{PA}} \Box_{\mathsf{PA},m} C$. Ergo, since $\prec_i$ preserves Σ-sentences —remember that for $i = 1$ this uses Matiyasevič's Theorem—, $\mathcal{N} \models \Box_{\mathsf{PA},m} C$. Ergo, by $\mathsf{REF}_{\mathcal{M}}$, $\mathcal{N} \models C$. Conversely, suppose

$$\forall C{\in}\mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}\ (\mathcal{M} \models \Box_{\mathsf{PA}} C \Rightarrow \mathcal{N} \models C).$$

By verifiable uniform essential reflexiveness, we have that, for any m in $\mathcal{M}$ and for any D in $\mathcal{L}^{\mathcal{M}}_{\mathsf{PA}}$, $\mathcal{M} \models \Box_{\mathsf{PA}}(\Box_{\mathsf{PA},m} D \to D)$. Ergo $\mathcal{N} \models \Box_{\mathsf{PA},m} D \to D$. ❑

We will consider the models of PA equiped with accessibility relations $\prec_{i+}, \preceq_i$, for a fixed i, in the roles of R, S, as a simplified Veltman frame —dropping the demand of upwards wellfoundedness of R. Clearly this frame is a class. However, I do not think that this matter needs to bother us here. If the reader wishes, she can restrict the big frame e.g. to countable models coded in, say, the standard natural numbers.

At this point it is convenient to change our language. We will use $a, b, c, \ldots$ for models of PA, $a \Vdash A$ for $a \models A$. We fix an i and write R, S for $\prec_{i+}, \preceq_i$.

Let $\mathcal{L}^{+}_{\mathsf{PA}}$ be the smallest language containing $\mathcal{L}_{\mathsf{PA}}$, closed under the logical connectives of Predicate Logic and under the connectives $\Box$ and $\rhd$ of Interpretability Logic. Define $\mathcal{L}^{a,+}_{\mathsf{PA}}$ in a similar way. Define, for A in $\mathcal{L}^{a,+}_{\mathsf{PA}}$, $a \Vdash \Box A$ and $a \Vdash A \rhd B$ in the usual way of forcing in (simplified) Veltman models. Viewed in this way, our big frame with $\Vdash$ is a big model, which we will call Big (or, more precisely, Big_i).

Theorem D.3 *Let A, B be sentences of $\mathcal{L}^{a}_{\mathsf{PA}}$. We have:*

$$a \Vdash A \rhd_{\mathsf{PA}} B \ \Leftrightarrow\ a \Vdash A \rhd B$$

Note that the similar, simpler result for $\Box$ is an immediate consequence of the result for $\rhd$.

Proof

"⇒" Suppose $a \Vdash A \rhd_{\mathsf{PA}} B$ and $aRb \Vdash A$. By the Orey-Hájek characterization, $a \Vdash \forall x \Box_{\mathsf{PA}}(A \to \Diamond_{\mathsf{PA},x} B)$. Moreover, for all $C \in \mathcal{L}^a_{\mathsf{PA}}$ such that $a \Vdash \Box_{\mathsf{PA}} C$, we have: $b \Vdash C$. So, for all k in a, $b \Vdash \Diamond_{\mathsf{PA},k} B$. We want to find a c with $bSc \Vdash B$ and aRc. Construct inside b, a restricted internal model c for $\mathsf{PA} + B$ by the formalized Henkin construction, using the Feferman Predicate $\triangle_{\mathsf{PA},B}$. This Feferman predicate is defined as follows:

$$\triangle_{\mathsf{PA},B} C :\leftrightarrow \exists x(\Box_{\mathsf{PA}+B,x} C \wedge \Diamond_{\mathsf{PA},x} B).$$

It is clear that $bSc \models B$. By the transitivity of S: aSc. Consider C in $\mathcal{L}^a_{\mathsf{PA}}$ and k in a. Evidently, for some k' in a, $b \Vdash \Box_{\mathsf{PA},k'}(\Box_{\mathsf{PA},k} C \to C)$. Hence, $b \Vdash \triangle_{\mathsf{PA},B}(\Box_{\mathsf{PA},k} C \to C)$. We may conclude that $c \Vdash \Box_{\mathsf{PA},k} C \to C$ and, thus, aRc.

"⇐" Suppose $a \nVdash A \rhd_{\mathsf{PA}} B$. Then, by Orey-Hájek, for some k in a, $a \Vdash \Diamond_{\mathsf{PA}}(A \wedge \Box_{\mathsf{PA},k} \neg B)$. Using the Henkin construction, we may build b, with $aRb \Vdash A$ and $b \Vdash \Box_{\mathsf{PA},k} \neg B$. Consider any c with $bS_a c$. Since bSc, it follows that $c \Vdash \Box_{\mathsf{PA},k} \neg B$. Since aRc, it follows that $c \Vdash \neg B$. ❑

Clearly theorem D.3 allows us to translate modulo valid equivalence the arithmetico-modal language back into its purely arithmetical fragment.

We want to compare ILM-models with Big. Since we are considering models of different signature the usual notion of bisimulation will not do. The following minor adaptation will do the trick. Consider two languages $\mathcal{U}$ and $\mathcal{V}$ that are both closed under the connectives of interpretability logic. Consider finite sets of sentences $X \subseteq \mathcal{U}$ and $Y \subseteq \mathcal{V}$. Let $\mathcal{L}_{\mathsf{IL}}(X)$ be the sublanguage of $\mathcal{U}$ generated by X and the connectives of interpretability logic. Similarly for $\mathcal{L}_{\mathsf{IL}}(Y)$ and $\mathcal{V}$. Let σ be a total and surjective relation between X and Y. We extend σ to the smallest relation σ' between $\mathcal{L}_{\mathsf{IL}}(X)$ and $\mathcal{L}_{\mathsf{IL}}(Y)$ closed under rules like:

- if $A\sigma' B$ and $A'\sigma' B'$ then $(A \wedge A')\sigma(B \wedge B')$.

etcetera. We rename σ', par abus de langage, to σ. Note that σ is again total and surjective between $\mathcal{L}_{\mathsf{IL}}(X)$ and $\mathcal{L}_{\mathsf{IL}}(Y)$. Moreover if σ is functional on X, then σ is functional on $\mathcal{L}_{\mathsf{IL}}(X)$. In this case, we write $A\sigma$ for the unique B with $A\sigma B$. Let $\mathcal{K}$ and $\mathcal{M}$ be models for $\mathcal{U}$, respectively $\mathcal{V}$, with a forcing relation that is 'correct' w.r.t. the connectives of interpretability logic. A relation $\mathcal{B}$ between the nodes of $\mathcal{K}$ and $\mathcal{M}$ is a *σ-bisimulation* if:

- For all $A \in X$, $B \in Y$: $(k\mathcal{B}m \text{ and } A\sigma B) \Rightarrow (k \Vdash A \Leftrightarrow m \Vdash B)$
- $(k\mathcal{B}m$ and $kRk') \Rightarrow \exists m'(k'\mathcal{B}m', mRm'$ and
  $\forall m''\ (m'S_m m'' \Rightarrow \exists k''(k''\mathcal{B}m''$ and $k'S_k k'')))$
- $(k\mathcal{B}m$ and $mRm') \Rightarrow \exists k'(k'\mathcal{B}m',\ kRk'$ and
  $\forall k''(k'S_k k'' \Rightarrow \exists m''(k''\mathcal{B}m''$ and $m'S_m m'')))$

It is easily seen that for any A, B in $\mathcal{L}_{\mathsf{IL}}(X)$, respectively $\mathcal{L}_{\mathsf{IL}}(Y)$, with $A\sigma B$ we have: $k\mathcal{B}m \Rightarrow (k \Vdash A \Leftrightarrow m \Vdash B)$.

Consider any finite ILM-model $\mathcal{K}$, for a finite set of atoms $\mathcal{P}$, with domain $\{1, \cdots, N\}$ and bottom 1. By [62], or by [19], we can find arithmetical sentences λ_i, for $i = 1, \cdots, N$, such that for all $k, m, n \in \{1, \cdots N\}$:

Z0 $m \neq n \Rightarrow \mathsf{PA} \vdash \neg(\lambda_m \wedge \lambda_n)$

Z1 $\mathsf{PA} + \lambda_m$ is consistent

Z2 $\mathsf{PA} \vdash \lambda_k \rightarrow \Box \bigvee_{kRm} \lambda_m$

Z3 $mS_kn \Rightarrow \mathsf{PA} \vdash \lambda_k \rightarrow \lambda_m \rhd_{\mathsf{PA}} \lambda_n$

Z4 $kRm \Rightarrow \mathsf{PA} \vdash \lambda_k \rightarrow \neg(\lambda_m \rhd_{\mathsf{PA}} \neg \bigvee_{mS_kn} \lambda_n)$

(In Z1 it would suffice to ask that $\mathsf{PA} + \lambda_1$ is consistent, since the other consistencies follow from this and Z4.) Define $p\lambda := \bigvee_{m \Vdash p} \lambda_m$. We take $\mathcal{P}$ in the role of X and the range of λ in the role of Y. Define the following relation between the nodes of $\mathcal{K}$ and the nodes of Big: $k\Lambda a :\Leftrightarrow a \Vdash \lambda_k$. We have:

Theorem D.4 Λ *is a total* λ*-bisimulation.*

Proof

Totality is immediate from Z1. Suppose $k\Lambda a$.

If $k \Vdash p$, then, by definition, $a \Vdash p\lambda$. Conversely, if $a \Vdash p\lambda$, then $a \Vdash \lambda_m$, for some m with $m \Vdash p$. Since, $a \Vdash \lambda_k$, we have, by Z0, $m = k$ and, hence, $k \Vdash p$.

Suppose kRm. By Z4, $a \Vdash \neg(\lambda_m \rhd_{\mathsf{PA}} \neg \bigvee_{mS_kn} \lambda_n)$. By theorem D.3, there is a b, such that aRb, $b \Vdash \lambda_m$ and such that, for all c with bS_ac, $c \Vdash \bigvee_{mS_kn} \lambda_n$. In other words, $m\Lambda b$ and, whenever bS_ac, then $n\Lambda c$, for some n with mS_kn.

Suppose aRb. By Z2 and theorem D.3, there is an m, with kRm and $b \Vdash \lambda_m$, i.e. $m\Lambda b$. Consider any n with mS_kn. Since, by Z3, $a \Vdash \lambda_m \rhd_{\mathsf{PA}} \lambda_n$, we can find, by theorem D.3, a c with bS_ac and $c \Vdash \lambda_n$, in other words, $n\Lambda c$. ❑

Note that —modulo our switches between $\rhd_{\mathsf{PA}}$ and $\rhd$— there is nothing arithmetical about the proof of D.4. An immediate consequence is the arithmetical completeness of ILM. Let σ be a function from $\mathcal{L}_{\mathsf{IL}}$ to the sentences of $\mathcal{L}_{\mathsf{PA}}$. We write $\phi\sigma$ as before and ϕ^σ for the usual arithmetical interpretation of ϕ corresponding to σ. Note that ϕ^σ can be obtained from ϕ^σ by 'replacing' $\rhd$ by $\rhd_{\mathsf{PA}}$ and $\Box$ by $\Box_{\mathsf{PA}}$. We have, by theorem D.3: $\mathsf{Big} \Vdash \phi\sigma \Leftrightarrow \mathsf{Big} \Vdash \phi^\sigma$.

Theorem D.5 $\mathsf{ILM} \vdash \phi \Leftrightarrow \forall\sigma\ \mathsf{Big} \Vdash \phi\sigma \Leftrightarrow \forall\sigma\ \mathsf{PA} \vdash \phi^\sigma$

Proof

The second equivalence is by D.3 and the completeness theorem for predicate Logic. The first equivalence from left to right is the usual check of the soundnes of ILM. We prove the second equivalence from right to left by contraposition. Suppose ILM $\nvdash \phi$, then, by [12], there is a finite ILM-countermodel $\mathcal{K}$ to ϕ. We may assume that the domain of $\mathcal{K}$ is $\{1, \cdots, N\}$, that 1 is the root of $\mathcal{K}$ and that $1 \nVdash \phi$. Consider λ for our present model. By Theorem D.4 there is a PA-model a that λ-bisimulates with 1. We have: $a \nVdash \phi\sigma$. ❑

We end this appendix, by considering a model connective on Big that cannot be eliminated via an arithmetical definition. Let $\mathcal{L}^{*}_{\mathsf{PA}}$ be the language of arithmetic extended by a new unary connective $\heartsuit$. The analogous language with parameters in a, is $\mathcal{L}^{a,*}_{\mathsf{PA}}$. Define, for $A \in \mathcal{L}^{a,*}_{\mathsf{PA}}$,

- $a \Vdash \heartsuit A \;:\Leftrightarrow\; \forall b (aSb \Rightarrow b \Vdash A)$.

By theorem D.1, for arithmetical A the forcing of $\heartsuit A$ is independent of the question which $\preceq_i$ we have chosen S to be.

Let $\mathcal{L}_{\mathsf{ML}}$ be a language of ordinary unimodal logic on finitely many variables. We take the single modal necessity operator to be $\heartsuit$. Let σ map the variables to arithmetical sentences. We extend this mapping to $\mathcal{L}_{\mathsf{ML}}$ in the obvious way. Define $\Theta := \{\phi \in \mathcal{L}_{\mathsf{ML}} \mid \forall \sigma \; \mathsf{Big} \models \phi\sigma\}$.

Theorem D.6 Θ *is closed under the* S4*-axioms and -rules.*

Proof

The theorem is immediate from the fact that S is a weak partial ordering. ❑

We will show in theorem D.9 that Θ coincides with S4.

Corollary D.7 There is no arithmetical formula α such that, for all arithmetical sentences A, $a \Vdash (\heartsuit A \leftrightarrow \alpha(\underline{\#A}))$. ❑

Proof

The non-existence of α is immediate from the well-known fact that S4 plus modalized self-reference yields inconsistency. We can certainly afford the space to repeat the argument. Let's agree to write αA for $\alpha(\underline{\#A})$. Find, by the Gödel Fixed Point Lemma, G with $\mathsf{PA} \vdash G \leftrightarrow \neg\alpha G$. Suppose $a \Vdash \alpha G$, then $a \Vdash G$ and, hence $a \Vdash \neg\alpha G$. Quod non. Hence, for no a, $a \Vdash \alpha G$. On the other hand, if $a \Vdash \neg\alpha G$, then there is a b with aSb and $b \Vdash \neg G$ and thus $b \Vdash \alpha G$. A contradiction. ❑

Not only is $\heartsuit$ arithmetically undefinable, it falsifies induction for $\mathcal{L}^*_{\mathsf{PA}}$ in Big, since we can use $\heartsuit$ to define the (standard) natural numbers, as the following theorem shows.

Theorem D.8 *There is a $\mathcal{L}^*_{\mathsf{PA}}$-fromula in one variable that defines the (standard) natural numbers in a. As a consequence true arithmetic can be interpreted into $\{A{\in}\mathcal{L}^*_{\mathsf{PA}} \mid \mathsf{Big} \not\Vdash A\}$.*

Proof

We show that the predicate $\heartsuit\Diamond_{\mathsf{PA},x}\top$ defines the natural numbers in a. Say $X := \{m{\in}a \mid a \Vdash \heartsuit\Diamond_{\mathsf{PA},\underline{m}}\top\}$. Since, for every $n \in \omega$, $\mathsf{PA} \vdash \Diamond_{\mathsf{PA},\underline{n}}\top$, we have: $\omega \subseteq X$. Consider a *non-standard* element m of a. It is clearly sufficient to show that $m \notin X$. In case $a \Vdash \Box_{\mathsf{PA},m}\bot$, we are done (by the identity interpretation). In case $a \Vdash \Diamond_{\mathsf{PA},m}\top$, we have, by the Second Incompleteness Theorem, $a \models \Diamond_{\mathsf{PA},m}\Box_{\mathsf{PA},m}\bot$. We use the formalized Henkin Construction to build (in a) an internal model b of $\Box_{\mathsf{PA},m}\bot$. This internal model is the desired witness that $m \notin X$. ❑

The following theorem is due to Volodya Shavrukov. It is published here by his permission.

Theorem D.9 $\mathsf{S4}$ *is the schematic modal logic of $\heartsuit$. In other words, Θ is precisely the set of theorems of* $\mathsf{S4}$.

Proof

Suppose ϕ is unprovable in $\mathsf{S4}$. We have a finite, transitive, reflexive Kripke model $\mathcal{K} = \langle K, S, \Vdash\rangle$ and k in $\mathcal{K}$ s.t. $k \not\Vdash \phi$.

Add an R-bottom node 0 under $\mathcal{K}$ to obtain a simplified ILM-model. Call the new model $\mathcal{K}^+$. Let k, m, n range over the domain K^+ of $\mathcal{K}^+$. Now we apply the Berarducci-Japaridze conditions for finite simplified ILM-models (see [5],[19]), extended with an extra insight that is immediate from the proof.[21] There is a function $\lambda : k \mapsto \lambda_k$ from K^+ to sentences of arithmetic, with the following properties.

1. $\mathsf{PA} \vdash \bigvee_{k\in K^+} \lambda_k$
2. $\sigma_k := \bigvee_{\{m\in K^+ \mid kSm\}} \lambda_m$ is Σ_1
3. $m \neq n \Rightarrow \mathsf{PA} \vdash \neg(\lambda_m \wedge \lambda_n)$
4. $k \neq 0 \Rightarrow \mathsf{PA} \vdash \lambda_k \to \Box \bigvee_{kRm} \lambda_m$

[21] Berarducci and Japaridze assume also that the original model $\mathcal{K}$ has an R-bottom. This assumption is never used in the proof.

5. $mS_kn \Rightarrow \mathsf{PA} \vdash \lambda_k \to \lambda_m \rhd_{\mathsf{PA}} \lambda_n$

6. $kRm \Rightarrow \mathsf{PA} \vdash \lambda_k \to \neg(\lambda_m \rhd_{\mathsf{PA}} \neg \bigvee_{mS_kn} \lambda_n)$

7. λ_0 is true

We claim: $kS_0m \Leftrightarrow \mathbb{N} \models \lambda_k \rhd_{\mathsf{PA}} \lambda_m$. From left to right is immediate by 5,7. For the converse direction, suppose that not kS_0m. then, by 6, we have $\mathbb{N} \models \neg(\lambda_k \rhd_{\mathsf{PA}} \neg \bigvee_{kS_0n} \lambda_n)$. But this is incompatible with $\mathbb{N} \models \lambda_k \rhd_{\mathsf{PA}} \lambda_m$, by 3. Hence $\mathbb{N} \not\models \lambda_k \rhd_{\mathsf{PA}} \lambda m$.

We define $k\Lambda a :\Leftrightarrow a \Vdash \lambda_k$. We show that Λ is a total bisimulation for ordinary modal logic, w.r.t. λ defined on the atoms as before, between $\mathcal{K}$ and Big. Totality is trivial. We leave the atomic case to the reader. Suppose $k \in K$, $k\Lambda aSb$. By 2 and the properties of S, we have $b \vdash \bigvee_{\{m \in K^+ \mid kSm\}} \lambda_m$, and, hence, for some m, $kSm\Lambda b$. Suppose $k \in K$, $k\Lambda a$ and kSm. By the choice of our model, it follows that kS_0m, and, hence, $\mathbb{N} \models \lambda_k \rhd_{\mathsf{PA}} \lambda_m$. Since $a \Vdash \lambda_k$, the interpretation of $\mathsf{PA} + \lambda_m$ in $\mathsf{PA} + \lambda_k$ gives us a b, such that $aSb \Vdash \lambda_m$ and, ipso facto, $m\Lambda b$. ❑