The Interpretability Logic Of *All* Reasonable Arithmetical Theories

The New Conjecture

Joost J. Joosten (joost.joosten@phil.uu.nl) and Albert Visser (albert.visser@phil.uu.nl)
Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht

Abstract. This paper is a presentation of a status quæstionis, to wit of the problem of the interpretability logic of *all* reasonable arithmetical theories. We present both the arithmetical side and the modal side of the question.

Keywords: arithmetic, interpretations, modal logic

MSC2000 codes: 03B45, 03F25, 03F30, 03F45

Dedicated to Dick de Jongh on the occasion of his 60th birthday

Table of Contents

1. Introduction

What challenges does the future have in store for us? When talking provability and interpretability logic, we are in the happy position of being able to give a pretty definite answer. Three great problems is what we are facing. The first —studied by R. Verbrugge and A. Berarducci, see (Verbrugge, 1993), (Berarducci and Verbrugge, 1993)— is the problem of the provability logics of Buss' S^1_2 and Wilkie & Paris' $I\Delta_0 + \Omega_1$. The second is the problem of the provabilitity logic of Heyting's Arithmetic —studied by A. Visser and R. Iemhoff, see (Visser, 1985), (Visser, 1994), (Visser et al., 1995), (Visser, 1998b), (Iemhoff, 20XX)). The third problem, the problem explained in this paper, is the problem of the interpretability logic of all reasonable arithmetical theories.

In this article, the current status of the problem will be presented. The paper provides the necessary definitions and a detailed explanation of the latest conjecture. It will be made evident that the problem is a good problem in that it intertwines modal and arithmetical ideas.

We did our best to make this exposition accessible to all readers with a modicum of mathematical sophistication. The next subsection is a brief introduction to interpretations.

1.1. What is an Interpretation?

The interpretations we are interested in are *relative interpretations* in the sense of Tarski, Mostwoski and Robinson (see (Tarski et al., 1953)). Consider theories U with language $\mathcal{L}_U$ and T with language $\mathcal{L}_T$. For the moment we assume that $\mathcal{L}_U$ is a relational language. An interpretation $\mathcal{K}$ of U in T is given by a pair $\langle \delta(x), F \rangle$. Here $\delta(x)$ is an $\mathcal{L}_T$-formula representing the *domain* of the interpretation.[1] F is a mapping that associates to each relation symbol R of $\mathcal{L}_U$ with arity n an $\mathcal{L}_T$-formula $F(R)(x_1, \cdots, x_n)$. Here $x_1, \ldots, x_n$ are suitably chosen free variables. We translate the formulas of $\mathcal{L}_U$ to the formulas of $\mathcal{L}_T$ as follows:

- $\mathcal{K}(R(y_1, \cdots, y_n)) := F(R)(y_1, \cdots, y_n)$,
  (We do not demand that identity is translated as identity.)
- $\mathcal{K}$ commutes with the propositional connectives,
- $\mathcal{K}(\forall y\, A) := \forall y\, (\delta(y) \to \mathcal{K}(A))$,
- $\mathcal{K}(\exists y\, A) := \exists y\, (\delta(y) \wedge \mathcal{K}(A))$,

There are some trifling details —e.g. about avoiding variable clashes— that we ignore here. In case $\mathcal{L}_U$ contains functionsymbols, we first apply

the usual algorithm to eliminate functionsymbols to translate $\mathcal{L}_U$ to a corresponding relational language and *then* we apply the translation sketched above. (For an attempt to get all the details right, see (Visser, 1998a).) Finally, we demand of interpretations that for all sentences A which are universal closures of axioms of U, we have $T \vdash \mathcal{K}(A)$.

We will write $\mathcal{K} : T \rhd U$ for $\mathcal{K}$ *is an interpretation of* U *in* T. An alternative notation, which is more suitable if we want to study the category of interpretations, is $U \xrightarrow{\mathcal{K}} T$. We write $T \rhd U$, for $\mathcal{K} : T \rhd U$, *for some* $\mathcal{K}$.

Interpretations are used for various purposes: to prove relative consistency, conservation results and undecidability results. The syntactical character of interpretations has the obvious advantage that it allows us to convert proofs of the interpreted theory in an efficient way into proofs of the interpreting theory. Examples of relative interpretations are e.g. the interpretation of arithmetic in set-theory, the interpretation of elementary syntax in arithmetic, the interpretation of $\mathsf{PA}+\mathsf{incon}(\mathsf{PA})$ in PA.

Let's forget, for a brief moment, about interpretations. Let's think about e.g. the construction of a model of two dimensional elliptic space in Euclidean three dimensional space. This is a construction inside the standard model of Euclidean geometry, which is modulo isomorphism the unique model of the second order version of three dimensional Euclidean geometry, of (modulo isomorphism) the standard model of two dimensional elliptic geometry. We construct this model by stipulating that 'point' in the new sense will be *line through a given point*[2], 'line' in the new sense is *plane through the given point*, 'incidence of point and line' is *the line representing the point is in the plane representing the line*, etc. If you inspect the construction, you will see that it just uses the resources of the first order theory of three dimensional Euclidean geometry. Thus it provides a uniform way of transforming models of (first order) three dimensional Euclidean geometry into models of (first order) two dimensional elliptic geometry. A still closer inspection shows that our construction can be viewed as a purely syntactical transformation. It provides a relative interpretation of (first order) two dimensional elliptic geometry in (first order) three dimensional Euclidean geometry.

We can capture the relation of interpretations and model constructions as follows. Let $\mathfrak{Mod}(V)$ be the class of models of a theory V. An interpretation $\mathcal{K}$ of U in T provides a uniform way to build internal models of U inside models of T. Thus $\mathcal{K}$ provides us with a function, say $\mathfrak{Mod}(\mathcal{K})$, from $\mathfrak{Mod}(T)$ to $\mathfrak{Mod}(U)$. Thus defined $\mathfrak{Mod}$ is a contravariant functor from $\mathfrak{Theory}$, the category of theories and interpretations, to $\mathfrak{Class}$, the category of definable classes and definable functions between classes. The idea of an interpretation as an 'internal

model given in a uniform way' is an important heuristic in thinking about interpretations: the mind craves reality and visualisation rather than syntax. We will exploit this heuristic in what follows. E.g. we will speak about one interpretation $\mathcal{K}$ being an end-extension of another one $\mathcal{M}$, meaning that in every model the internal model associated to $\mathcal{K}$ is an end-extension of the internal model of $\mathcal{M}$ in a uniform way.

1.2. So What's Reasonable?

We will be interested in interpretability of reasonable arithmetical theories. More specifically we will be interested in what such theories have to say about interpretability in these theories themselves. So what are reasonable arithmetical theories?

A theory, for our present purposes, is a predicate logical theory axiomatized by axioms in an axiom set that is given by arithmetical formula α. Unless stated otherwise we assume that α is simple, say a predicate corresponding with a class that is decidable by a p-time algorithm. Note that our specification makes *theory* an intensional notion, since we consider theories also from the point of view of theories: α and β may specify the same axiom set, but a theory U thinking about T_α and T_β need not be aware of that.

'Arithmetical' has a primary and a secondary meaning. In the primary meaning an arithmetical theory is an extension of Robinson's Arithmetic Q in the usual language of arithmetic with $\underline{0}$, S, $+$ and $\times$. (We will often use $\cdot$ instead of $\times$.) In the secondary meaning, an arithmetical theory is a pair $\langle T, \mathcal{N}\rangle$, where $\mathcal{N} : T \rhd \mathsf{Q}$. In other words an arithmetical theory is a theory with designated predicates representing the natural numbers, representing zero, etc. e.g. $\langle \mathsf{ZF}, \mathcal{N}\rangle$, where $\mathcal{N}$ is the usual set-theoretical representation of the natural numbers, is an arithmetical theory. The intended meaning of arithmetical theory in this paper is the secondary one.

The caution concerning the explicit designation of the natural numbers is necessary, since not all interpretations of number theory are provably isomorphic in a given theory. Thus the following three statements are equally true.

1. $\mathsf{con}(\mathsf{ZF})$ is independent of GB (= Gödel-Bernays Set Theory).
2. GB proves $\mathsf{con}(\mathsf{ZF})$.
3. GB proves $\neg\, \mathsf{con}(\mathsf{ZF})$.

Here $\mathsf{con}(\mathsf{ZF})$ abbreviates a fixed arithmetical sentence, but we vary, in the examples, the designated set of natural numbers. The correct formulation of our statements is:

1. $\mathsf{con}(\mathsf{ZF})$ is independent of $\langle \mathsf{GB}, \mathcal{N} \rangle$.

2. $\langle \mathsf{GB}, \mathcal{I} \rangle$ proves $\mathsf{con}(\mathsf{ZF})$.

3. $\langle \mathsf{GB}, \mathcal{K} \rangle$ proves $\neg\, \mathsf{con}(\mathsf{ZF})$.

Here $\mathcal{N}$ is the usual interpretation of the natural numbers in ZF lifted to GB. $\mathcal{I}$ is a definable cut of the $\mathcal{N}$-numbers and $\mathcal{K}$ is a suitable interpretation built using a syntactic variant of the Henkin construction. Both $\mathcal{N}$ and $\mathcal{I}$ are standard in that they represent the ordinary natural numbers (modulo isomorphism) in the standard model. $\mathcal{K}$, of course, cannot represent the standard numbers inside any model.

Rather than viewing the possibility of having different sets of numbers as a nuisance, we will make grateful use of it by switching between choices of what 'the numbers' are.

There is some arbitrariness in our singling out arithmetic as the thing we are interested in, especially since representations of syntax play such a large role in Gödelean metamathematics. We could as have well decided to speak about *syntactical* theories, counting e.g. $\langle \mathsf{PA}, \mathcal{S} \rangle$ as a syntactical theory, where $\mathcal{S}$ is a designated interpretation of some reasonable theory of elementary syntax. There are two reasons we will make the traditional choice to speak about arithmetic: first simply because it is the traditional choice —changing it will cause confusion—, secondly the methodology of definable cuts is easier to understand in the context of arithmetic.

What is reasonable? It means *at least*: strong enough to verify the minimal principles we are interested in. Take, e.g. the principle that tells us that if something is provable, then it's provable that it's provable. To verify the principle in the obvious way, we need Δ_0-induction, plus the totality of the function ω_1, where $\omega_1(x) = 2^{(\log_2 x)^2}$. This principle is called Ω_1. Another principle is the one stating that interpretations can be composed, i.e. if $\mathcal{K} : U \rhd V$ and $\mathcal{M} : V \rhd W$, then $\mathcal{M} \circ \mathcal{K} : U \rhd W$. To verify it we need Σ-collection, also known as $B\Sigma_1$, the principle $\forall x \leq a\, \exists y\, A \to \exists b \forall x \leq a\, \exists y \leq b\, A$, where A is Σ_1.

Thus we demand that reasonable arithmetical theories contain a minimal arithmetical theory Basic. Formally: a reasonable arithmetic is of the form $\langle T, \mathcal{N} \rangle$, where $\mathcal{N} : T \rhd \mathsf{Basic}$. The most plausible choice for Basic at the time of writing is $I\Delta_0 + \Omega_1 + B\Sigma_1$. (See (Visser, 1991), for some shameless trickery to get rid of the assumption $B\Sigma_1$.)

The second demand that we pose, has to do with the coherence of the theories. Given two theories T, U, we could take a disjoint union $T \oplus U$ in such a way that the T- and the U-objects have no recognizable interaction at all. So if our numbers are 'confined in' T, they will not be able to 'interact' in any way with the U-part. What we demand

is that our theory is sequential: it should contain (in the sense of interpretability) a theory of sequences of *all* objects the theory can talk about. Here the lengths $\mathsf{length}(\sigma)$ and projections $(\sigma)_i$ are taken from an initial segment of the designated numbers. Sequentiality is important to make e.g. the construction of partial truthpredicates possible. For more on sequentiality, see e.g. (Hájek and Pudlák, 1991).

The third demand is not really a demand but a programmatic point. We should keep the answer to the question what a reasonable theory is, somewhat indefinite. The class of reasonable theories is that class of theories that allows a beautiful answer to the question what the interpretability logic of all reasonable theories is. E.g. it could happen that only the theories that contain the axiom *that exponentiation is total* have a nice logic. Well, in that case we say that those theories are the reasonable ones.

1.3. Approaches to Interpretability

What could the metamathematical study of interpretability and interpretations look like? One idea is to study *degrees of interpretability*. Interpretability yields a partial preorder on theories. Dividing the associated equivalence relation out we get a degree-theory. Degree-theory has been studied by P. Lindström and C. Bennet (see e.g. (Lindström, 1997) and (Bennet, 1986)) and by V. Švejdar (see (Švejdar, 1978)). The work on degrees was very fruitful as a generator of methods and techniques. Some of these techniques have been adapted for application in interpretability logic.

We feel that it could be very fruitful to extend the degree-theoretic approach to the study of suitable categories of interpretations. The more expressive category-theoretical language might be better suited to express certain basic insights concerning interpretability. There were some attempts to initiate such a study, but these attempts did yield less than satisfactory results. Some further experimentation is needed to isolate the right categories.

The approach to interpretability that is the focus of this paper is the modal study of interpretability.[3] The modal language has the advantage of expressiveness, but there are costs. First modal logic is about 'propositions' not about theories. This means that we cannot directly study the relations $\mathcal{K} : T \rhd U$ or even $T \rhd U$. What we study is the relation $A \rhd_T B$, which is defined as follows:

- $A \rhd_T B \;\; :\Leftrightarrow (T + A) \rhd (T + B)$.

Here T is the base theory. We speak of (sentential) interpretability *over* T. Secondly, we are interested in iterating the modal connectives. We

want to allow things like $(A \rhd B) \rhd C$. This means that our research is restricted to base theories T that have sufficient coding ability to formalize a decent amount of reasoning concerning interpretations. This restriction is substantial since lots of important interpretations fall outside the scope of our investigation. If we pay the costs, there are some gains.

1. We have a modal language that is rich enough to articulate both the incompleteness theorems and the model existence lemma, which is the heart of the completeness theorem.
2. Some substantial reasoning concerning interpretability can be executed in this modal logic.
3. The Kripke model theory of the logic is highly interesting qua modal logic.
4. The arithmetical side of the study involves substantial arithmetical insights. As we will see, in an indirect way, our logic can talk about large and small numbers.

Before we introduce the modal logics, we interpolate a brief introduction to some salient arithmetical facts.

2. Parvulae Arithmeticae

2.1. Coding

Since the function $\omega_1 = \lambda x.2^{(\log_2 x)^2}$ is present in our basic system of arithmetic Basic, we have p-time computable functions available. Having these, arithmetization of syntax becomes a piece of cake. The most obvious gödelnumbering of strings in a given alphabet is also the best one. We enumerate first the strings of length 0, then the strings of length 1, and so on. The strings of the same length we order alphabetically. We assign to each string as gödelnumber simply its ordernumber in the sequence so obtained. It turns out, using a trick due to Smullyan, that operations on strings like concatenation can be easily arithmetically represented under this coding. An important insight is the elementary fact that the gödelnumber of a string σ is of order $A^{\mathsf{length}(\sigma)}$, where A is the cardinality of the alphabet. We will code syntactical elements, like formulas and proofs, by writing them out and then taking the code of the resulting string.

We will write $\Box_T A$ for the arithmetization of T *proves* A and $A \rhd_T B$ for the arithmetization of $T + A$ *interprets* $T + B$. If A contains a free

variable x, $\Box_T A$ is the arithmetization of *the result of substituting the numeral of x in A for "x" is provable in T'*. Further conventions are similar.

2.2. Efficient Numerals

It is definitely not a good idea to represent the number n by the numeral $\overbrace{S\cdots S}^{n}\underline{0}$. The gödelnumber of this numeral will be of order 2^{cn}, for a fixed constant c. So the function sending a number to the code of its numeral will be exponential. Exponentiation is not generally available in Basic. Hence we will use binary numerals instead. These are defined by $\mathsf{num}(0) := \underline{0}$, $\mathsf{num}(2n+1) := S(SS\underline{0} \cdot \mathsf{num}(n))$, $\mathsf{num}(2n+2) := SS\underline{0} \cdot \mathsf{num}(n+1)$. This representation has the happy consequence that the gödelnumber of the numeral of n is of order $2^{c \log_2 n}$, i.e. n^k, for some fixed standard k.

2.3. Numbers Large and Small

We have to face the basic fact that we are going to use theories that do not have full induction. Note that also quite strong theories may lack full induction, e.g. $\langle \mathsf{GB}, \mathcal{N} \rangle$. Thus, in our theories, it may happen that we have definable sets of numbers containing 0 and closed under successor such that the theory doesn't think this set contains all (designated) numbers. In some cases the theory will even positively know this set does not contain all numbers. Such definable sets of numbers, closed under successor but not necessarily containing all designated numbers, play an important role in the metamathematical study of arithmetics. For many purposes it is convenient to put stronger demands on these sets: we ask that they are *definable cuts*. Let $\langle T, \mathcal{N} \rangle$ be an arithmetical theory. Here $\mathcal{N} = \langle \delta, F \rangle$.

An $\mathcal{L}_T$-formula I is/presents a $\langle T, \mathcal{N} \rangle$-cut iff T proves that:

1. $Ix \to \delta x$,
2. $(Ix \wedge \mathcal{N}(x = y)) \to Iy$,
3. I is downwards closed under $<$, i.e.
   $(I(x) \wedge \mathcal{N}(y < x)) \to I(y)$,
4. I is closed under 0, S, $+$, $\times$ and ω_1, i.e.
   a) $\mathcal{N}(x = 0) \to Ix$,
   b) $(Ix \wedge \mathcal{N}(Sx = y)) \to Iy$,

c) $(Ix \wedge Iy \wedge \mathcal{N}(x + y = z)) \to Iz$,

d) $(Ix \wedge Iy \wedge \mathcal{N}(x \cdot y = z)) \to Iz$,

e) $(Ix \wedge \mathcal{N}(\omega_1 x = y)) \to Iy$.

Note that '$\omega_1 x = y$' is, in the usual set-up, an abbreviation of a complex formula.

We will sometimes write '$x \in I$' for 'Ix'.

Using a wonderful trick invented by Solovay (Solovay, 1976), we can always 'shorten' a definable set of numbers, T-provably closed under successor to a T-cut. Cuts can be considered as 'notions of smallness': the numbers inside the cut are 'small', the ones above it 'big'.

We will consider cuts themselves as interpretations of arithmetic, confusing the cut I with the interpretation $\langle I, F\rangle$, where F is the interpretation function associated with $\mathcal{N}$. It is easy to see that $\langle I, F\rangle$ is indeed an interpretation.

From this point on, we will often suppress the designated cut $\mathcal{N}$, writing as if $\mathcal{N}$ were the identity interpretation.

A startling fact about cuts is the *outside big, inside small* principle. Even if T may fail to believe that every number is in the T-cut I, we do have:

THEOREM 2.1. $T \vdash \forall x \, \Box_T \, x \in I$.

The point is that we can have big proofs showing that big numbers are small. Here is a somewhat more elaborate proofsketch.

PROOF OF 2.1. We reason informally in T. Let σ be a (standard) proof of $\forall x \; (x \in I \to S(SS\underline{0} \cdot x) \in I)$. We convert a proof π of $\underline{n} \in I$ into a proof of $S(SS\underline{0} \cdot \underline{n}) \in I$ as follows.

$$\dfrac{\pi \qquad \dfrac{\sigma}{\underline{n} \in I \to S(SS\underline{0} \cdot \underline{n}) \in I}\,\forall E}{S(SS\underline{0} \cdot \underline{n}) \in I}\,\to E$$

Similarly we convert a proof of $\underline{n} \in I$ into a proof of $(SS\underline{0} \cdot \underline{n}) \in I$. Clearly a proof of $\underline{n} \in I$ will use in the order of $\log_2 n$ steps. The number of symbols in a step of the proof can be estimated by $a \log_2 n + b$ for fixed standard numbers a and b. The number of symbols in the proof will be estimated by: $\log_2 n \cdot (a \log_2 n + b)$. Let $c := a + b$. We can replace our estimate by: $c \cdot (\log_2(n+2))^2$. So the size of the Gödelnumber of the proof will be estimated by $2^{c \cdot (\log_2(n+2))^2} = (\omega_1(n+2))^c$. The function $\lambda x.(\omega_1(x+2))^c$ is present in Basic. QED

2.4. Cuts and Interpretations

If we think of an interpretation $\mathcal{K}$ of U in T as an inner model of U inside a model of T, we can ask how the T-numbers do compare to the U-numbers as seen via $\mathcal{K}$. To be pedantically precise, if the T-numbers are given by $\mathcal{N}$ and if the U-numbers are given by $\mathcal{M}$, how does the internal model of Basic given by $\mathcal{N}$ compare to the internal model of Basic given by $\mathcal{K} \circ \mathcal{M}$?

In case T has full induction, the answer is simple. Let's for the moment step into the outside world and remind ourselves of a basic fact about non-standard models of arithmetic. The natural numbers form (modulo embedding) an initial fragment of every non-standard model. I.o.w. every non-standard model is an end-extension of the standard model. We can prove that fact by defining the embedding of the natural numbers into the non-standard model by external recursion and by subsequently proving the desired properties of the embedding by external induction. Essentially the same argument can be repeated in T. We define the embedding, using the fact that we are supposed to have sequences of objects, by an explicit T-formula and verify its properties with T-induction. The upshot is that the U-numbers as seen via $\mathcal{K}$ form an end-extension of the T-numbers.

Now what happens if T doesn't have full induction? The answer to this question has been provided by Pavel Pudlák in his fundamental paper (Pudlák, 1985). Here it is. There is a T-cut I such that the U-numbers as viewed via $\mathcal{K}$ are an end-extension of I.

The argument for Pudlák's theorem is a refinement of the usual argument sketched above: *where we lack induction, we compensate by switching to a smaller cut.*

A Closer Look

To convince the reader that the statement of Pudlák's theorem makes sense, we spell out the result in the pedantic mode. Remember that we assumed that T was really $\langle T, \mathcal{N} \rangle$ and U was really $\langle U, \mathcal{M} \rangle$. Now $\mathcal{Q} := \mathcal{K} \circ \mathcal{M}$ is an interpretation of Basic in T, representing the U-numbers as viewed by T via $\mathcal{K}$. Pudlák's Theorem tells us that there is a $\langle T, \mathcal{N} \rangle$-cut I such that there is, verifiably in T, a definable embedding of I into an initial segment of the U-numbers as viewed in T via $\mathcal{Q}$. This means that there is a T-formula E such that T proves:

1. $Exx' \to (Ix \wedge \delta_{\mathcal{Q}}(x'))$
   (E is a relation between I and $\delta_{\mathcal{Q}}$),
2. $(Exx' \wedge \mathcal{N}(x = y) \wedge \mathcal{Q}(x' = y')) \to Eyy'$
   (E is a congruence w.r.t. the relevant 'identities'),

3. $Ix \to \exists x'\, Exx'$
   (E is total on I),

4. $(Exx' \wedge Exy') \to \mathcal{Q}(x' = y')$
   (E is a function),

5. $(Exx' \wedge Eyx') \to \mathcal{N}(x = y)$
   (E is injective),

6. $(Exx' \wedge \mathcal{Q}(y' < x')) \to \exists y\, (\mathcal{N}(y < x) \wedge yEy')$
   (The E-image of I is downwards closed in $\mathcal{Q}$),

7. $(Exx' \wedge \mathcal{N}(x = 0) \to \mathcal{Q}(x' = 0)$
   (E commutes with 0),

8. $(Exx' \wedge Eyy' \wedge \mathcal{N}(Sx = y)) \to \mathcal{Q}(Sx' = y')$
   (E commutes with S),

9. $(Exx' \wedge Eyy' \wedge Ezz' \wedge \mathcal{N}(x + y = z)) \to \mathcal{Q}(x' + y' = z')$
   (E commutes with $+$),

10. $(Exx' \wedge Eyy' \wedge Ezz' \wedge \mathcal{N}(x \cdot y = z)) \to \mathcal{Q}(x' \cdot y' = z')$
    (E commutes with $\times$).

The image, say J, of the $\langle T, \mathcal{N}\rangle$-cut I is easily seen to be a $\langle T, \mathcal{Q}\rangle$-cut. T shows that E is an isomorphism between I and J. J is not generally internally definable in $\mathcal{K}$, i.o.w. there need not be an $\mathcal{L}_U$-formula G such that $T \vdash \exists x\, Exx' \leftrightarrow \mathcal{K}(Gx')$.

3. Interpretability Logic Explained

3.1. Description of the System IL

The language of interpretability logic, $\mathcal{L}_{\mathsf{int}}$, is the language of modal propositional logic extended with a binary modal operator $\rhd$. We read $A \rhd B$ as: A interprets B. We will write $\Diamond A$ as an abbreviation of $\neg\Box\neg$.

Let U be a reasonable arithmetical theory. 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:

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

We study the interpretability principles valid in theories U, asking ourselves for which C in the modal language we have $U \vdash C^*$, for all $(.)^*$ and asking ourselves which principles are valid in all reasonable

theories. The set of principles valid in U is called $\mathfrak{Il}(U)$. The set of principles valid in *all* reasonable theories will be called $\mathfrak{Il}(\mathsf{all})$.[4]

We introduce the basic modal logic IL. The principles of this logic are arithmetically sound for a wide class of theories and for various interpretations of its main connective $\rhd$.[5] The theory is arithmetically incomplete for all known arithmetical interpretations. The motivation for studying this specific set of axioms comes from its *modal* simplicity and elegance.

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 $\vdash A \Rightarrow \vdash \Box A$

L2 $\vdash \Box(A \to B) \to (\Box A \to \Box B)$

L3 $\vdash \Box A \to \Box\Box A$

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

J1 $\vdash \Box(A \to B) \to A \rhd B$

J2 $\vdash (A \rhd B \ \wedge \ B \rhd C) \to A \rhd C$

J3 $\vdash (A \rhd C \ \wedge \ B \rhd C) \to (A \vee B) \rhd C$

J4 $\vdash A \rhd B \to (\Diamond A \to \Diamond B)$

J5 $\vdash \Diamond A \rhd A$

L1-4 are the well-known principles of Löb's Logic. IL is certainly valid in all reasonable theories U. We will provide the arithmetical justifications of the principles in subsection 3.2.

D. de Jongh and A. Visser proved that IL has *unique and explicit fixed points*. See (de Jongh and Visser, 1991). No characterization of the *closed fragment* of IL has been given. IL satisfies *interpolation*, see (Areces et al., 1998). De Jongh and Veltman prove a *modal completeness theorem* w.r.t. Veltman models. See (de Jongh and Veltman, 1990).

Here is a sample of IL-reasoning. We prove: $\vdash A \rhd (A \wedge \Box\neg A)$. First, by L1-4, we can derive, taking the contraposition of L4: $\vdash \Diamond A \to \Diamond(A \wedge \Box\neg A)$. So, by L1 and J1, we find: $\vdash \Diamond A \rhd \Diamond(A \wedge \Box\neg A)$. Applying J5 and J2, we get: (a) $\vdash \Diamond A \rhd (A \wedge \Box\neg A)$. We also have, by L1 and J1: (b) $\vdash A \rhd ((A \wedge \Box\neg A) \vee \Diamond A)$ and (c): $\vdash (A \wedge \Box\neg A) \rhd (A \wedge \Box\neg A)$. Applying J3 and J2 to (a), (b) and (c) we arrive at the desired result.

Putting $\top$ for A in the principle we just derived, we see that it follows that one can construct, in a uniform way, inside every model of a given arithmetical theory T an internal model of $T + \mathsf{incon}(T)$.

3.2. The Arithmetical Validity of IL

Verification of the L*-principles*
It is well known that the principles of Löb's Logic can be derived in Buss' S_2^1 (see (Buss, 1986)) or in Wilkie & Paris' $I\Delta_0 + \Omega_1$ (see (Wilkie and Paris, 1987)). Since Basic is supposed to extend S_2^1, we are done. The proof of L3 is by induction on the subformulas of $\Box_T A$, using the fact that $\Box_T A$ is a $\exists\Delta_0^b$-predicate. A Δ_0^b-formula only has logarithmically bounded quantifiers.

It is a remarkable fact that L3 is doubly redundant in IL. By a clever argument, due to Dick de Jongh, we can derive L3 from L1,2,4. However this redundancy is not arithmetically helpful, since the usual arithmetical verification of Löb's axiom L4 uses the validity of L3.[6] The second way is to derive L3 from J5 and J4. The striking thing about this alternative proof is that it provides a really different way to obtain a T-proof of $\Box_T A$ from a T-proof of A.[7] L3 is one of those cases where we have one fact but two insights.

Both styles of proofs of L3, yield on inspection sharper results, like:

- $T \vdash \Box_T A \to \Box_T \Box_T^I A$
  Here I is any T-cut. We write $\Box_T^I A$ for $I(\Box_T A)$ —note that we need only to relativize the unbounded existential quantifier of $\Box_T A$ to I.)

- $T \vdash \Box_T A \to \Box_T \triangle_T A$
  Here $\triangle$ stands for either cutfree, Herband or tableaux provability.

The derivation of the first strengthening in the *induction on subformulas* style, runs as follows.

We obtain at a certain point $\Box_T \mathsf{proof}_T(x, \underline{\mathsf{gn}(A)})$. (Here gn is the gödelnumbering function.) The *outside big, inside small* principle tells us that $\Box_T x \in I$. Ergo, $\Box_T \Box_T^I A$.

Verification of J1
The validity of J1 is witnessed by the identity interpretation ID.

Verification of J2
If $\mathcal{K} : A \rhd_T B$ and $\mathcal{M} : B \rhd_T C$, then $(\mathcal{M} \circ \mathcal{K}) : A \rhd_T C$.

Verification of J3
J3 is valid, since, given any two interpretations $\mathcal{K}$ and $\mathcal{M}$ and any sentence A, we can construct an interpretation $\mathcal{K}[A]\mathcal{M}$, the disjoint A-sum of $\mathcal{K}$ and $\mathcal{M}$, that behaves like $\mathcal{K}$ if A and like $\mathcal{M}$ if $\neg A$. We take:

- $\delta_{\mathcal{K}[A]\mathcal{M}}(x) := ((\delta_{\mathcal{K}}(x) \wedge A) \vee (\delta_{\mathcal{M}}(x) \wedge \neg A))$,
- $(\mathcal{K}[A]\mathcal{M})(P)(\vec{x}) := ((\mathcal{K}(P)(\vec{x}) \wedge A) \vee (\mathcal{M}(P)(\vec{x}) \wedge \neg A))$

We find that if $\mathcal{K} : A \rhd_T C$ and $\mathcal{M} : B \rhd_T C$, then $(\mathcal{K}[A]\mathcal{M}) : (A \vee B) \rhd_T C$.

Verification of J4
J4 tells us that relative interpretability implies relative consistency. If we would have $\Box_T \neg B$ and $\mathcal{K} : A \rhd_T B$, then it would follow that $\Box_T(A \to \bot^{\mathcal{K}})$, and hence $\Box_T \neg A$.

Verification of J5
J5 is the *interpretation existence lemma*. It is the syntactical realization of the Henkin model existence lemma. Inspecting the usual proof of the model existence lemma, one sees that it involves the construction of a set of sentences describing a model. This set can as well be viewed as describing an interpretation. The set is constructed as a path in a binary tree. This path is described by a Δ^0_2-predicate. The desired properties of the set of sentences are verified using induction. Thus the whole argument can easily be verified in PA. The construction can be executed in almost any arithmetical theory by compensating for the lack of induction by switching to definable cuts.[8]

A moment's reflection shows that the choice of the numbers in which we execute the Henkin construction is irrelevant. So, in particular, this set of numbers, might very well be some T-cut I. Thus we arrive at the following sharpening of J5.

- $T \vdash \forall I\ (\Diamond^I_T A \rhd_T A)$.

Here $\Diamond^I_T A$ stands for $I(\Diamond_T A)$, which is T-equivalent to

$$\forall x \in I\ \neg\, \mathsf{prov}_T(x, \underline{\mathsf{gn}(\neg A)}),$$

where gn is the gödelnumbering function.

Note that the first strengthened version of L3, follows easily from J4 and the sharpened version of J5. We have $T \vdash \forall I\ (\Diamond^I_T B \rhd_T B)$, and, hence, $T \vdash \forall I\ (\Diamond_T \Diamond^I_T B \to \Diamond_T B)$. Replacing B by $\neg A$, contraposing and cleaning up spurious negations (using L1,2), we get: $T \vdash \forall I\ (\Box_T A \to \Box_T \Box^I_T A)$.

3.3. Beyond IL

IL is certainly arithmetically sound. However, it is not arithmetically complete for any reasonable arithmetical theory T and for any known interpretation of $\rhd$.

Montagna's Principle M

Let us first consider Peano Arithmetic, PA. The theory satisfies a further principle: Montagna's Principle.

$\mathsf{M} \quad \vdash A \rhd B \to (A \wedge \Box C) \rhd (B \wedge \Box C)$

The PA-validity of M was known independently to Švejdar and Lindström. Arithmetical completeness for the system ILM:=IL+M was conjectured by A. Visser. It was proved independently by V. Shavrukov (see (Shavrukov, 1988)) and A. Berarducci (see (Berarducci, 1990)). For nice presentations of the proof see also (Zambella, 1992) or (Japaridze and de Jongh, 1998). It turns out that ILM is sound and complete for all reasonable arithmetical theories satisfying full induction.[9] Here we just verify the arithmetical validity of M.

Let T have full induction. We prove the stronger principle:

– $\vdash A \rhd_T B \to (A \wedge S) \rhd_T (B \wedge S)$, for $S \in \Sigma^0_1$

Reason informally in T. Suppose $\mathcal{K} : (T + A) \rhd (T + B)$. Reason in T. (So we are in T two deep.) Suppose A and S. Consider $\mathcal{K}$. We will certainly have $\mathcal{K}(C)$ for each axiom C of T and $\mathcal{K}(B)$, since we have A. Now the $T + B$-numbers seen via $\mathcal{K}$ are an end-extension of the $T + A$-numbers, as explained in subsection 2.4. Moreover, Σ^0_1-sentences are preserved by end-extensions. Hence, we have $\mathcal{K}(S)$.

The Persistence Principle P

The persistence principle P is the following principle:

$\mathsf{P} \quad \vdash A \rhd B \to \Box(A \rhd B).$

The persistence principle is valid for interpretations in finitely axiomatized reasonable arithmetical theories T. The reason is simple. Let C be the conjunction of the axioms of T. Then, to obtain our principle, is is sufficient to verify:

$$T \vdash \exists \mathcal{K} \Box_T (A \to \mathcal{K}(C \wedge B)) \to \Box_T \exists \mathcal{K} \Box_T (A \to \mathcal{K}(C \wedge B))$$

which is obviously valid by verifiable $\exists \Delta^b_0$-completeness.

Albert Visser has shown that ILP is arithmetically complete for each finitely axiomatized reasonable theory that proves Superexp, the axiom stating that superexponentiation is total. See (Visser, 1990). It is definitely known that ILP is *not* complete for $I\Delta_0 + B\Sigma_1 + \mathsf{Exp}$, where Exp is the axiom stating that exponentiation is total.

The Principle W

The first principle that was found to be valid in all theories that strictly extends IL is the principle W. 'W' for "weak".

$\mathsf{W} \ \vdash A \rhd B \to A \rhd (B \wedge \Box \neg A)$

For some time it was conjectured that ILW was $\mathfrak{Il}(\mathsf{all})$. This conjecture was eventually refuted. Before we turn to the next principle, we verify W.

Remember our verification of the principle $\vdash A \rhd (A \wedge \Box \neg A)$. Now all the principles used remain valid if we relativize all the modal operators at nesting depth 1 to a T-cut I. Thus we obtain:

– $T \vdash \forall I \, (A \rhd_T (A \wedge \Box^I_T \neg A))$

Now reason in T. Suppose, for some $\mathcal{K}$, we have $\mathcal{K} : A \rhd_T B$. Let J be the $T + A$-cut such that, in $T + A$, J is (isomorphic to) an initial segment of the $T + B$-numbers viewed via $\mathcal{K}$. If J is not a T-cut, we may replace it by $J[A]\mathsf{ID}$, which certainly is a T-cut. (The notation $(\cdot)[\cdot](\cdot)$ was introduced when explaining $\mathsf{J3}$.) So we can assume that J is a T-cut. Now we have (a) $\mathcal{K} : (A \wedge \Box^J_T \neg A) \rhd_T (B \wedge \Box_T \neg A)$, since J is initial in $\mathcal{K}$ and since Σ^0_1-sentences are upwards persistent. By our previous consideration we have (b) $A \rhd_T (A \wedge \Box^J_T \neg A)$. Composing (a) and (b) we arrive at the desired result.

The Principle M_0
The next principle that was discovered is the principle M_0.

$\mathsf{M}_0 \ \vdash A \rhd B \to (\Diamond A \wedge \Box C) \rhd (B \wedge \Box C)$

Here is the arithmetical verification. Reason in T. Suppose $\mathcal{K} : A \rhd_T B$. Let J be the cut of the $T + A$-numbers, which is isomorphic to an initial segment of the $T + B$-numbers viewed via $\mathcal{K}$. As above, we may assume J to be a T-cut. We have $\Box_T(\Box_T C \to \Box_T \Box^J_T C)$ and, hence, $\Box_T((\Diamond_T A \wedge \Box_T C) \to \Diamond_T (A \wedge \Box^J_T C))$. It follows that:

$$\begin{aligned} \Diamond_T A \wedge \Box_T C \ &\rhd_T \ \Diamond_T (A \wedge \Box^J_T C) \\ &\rhd_T \ A \wedge \Box^J_T A \\ &\rhd_T \ B \wedge \Box_T C \end{aligned}$$

The last step, is witnessed by $\mathcal{K}$, using the fact that $\mathcal{K}$ is an end-extension of J and the upwards persistence of Σ^0_1-sentences.

M_0 can be viewed as a kind of 'M-ified' version of $\mathsf{J5}$. First note that we can rewrite $\mathsf{J5}$ as the equivalent:

– $\vdash A \rhd B \to \Diamond A \rhd B$

We get M_0, by plugging the $\Box C$'s into the consequent.

The Principle W^*
W^* is the following principle.

$$\mathsf{W}^* \vdash A \rhd B \to (B \wedge \Box C) \rhd (B \wedge \Box C \wedge \Box \neg A)$$

Dick de Jongh has shown that W^* is equivalent over IL with $\mathsf{W} + \mathsf{M}_0$.

The Principle P_0
For some time $\mathsf{ILWM_0}$ or, if you wish, ILW^* stood as the conjectured candidate for being $\mathfrak{Il}(\mathsf{all})$. Recently, Albert Visser found a new principle P_0.

$$\mathsf{P}_0 \vdash A \rhd \Diamond B \to \Box(A \rhd B)$$

The discovery of P_0 will be described in greater detail in subsection 4.4.

The arithmetical verification of P_0 is as follows. Reason in T. Suppose $\mathcal{K} : A \rhd_T \Diamond B$. Take a suitably large finite subtheory T_0 of T. (We can take T_0 standardly finite.) We certainly have $\mathcal{K} : (T + A) \rhd (T_0 + \Diamond_T B)$. Hence, (a) $\Box_T(\mathcal{K} : (T + A) \rhd (T_0 + \Diamond_T B))$. On the other hand, by interpretation-existence: (b) $\Box_T((T_0 + \Diamond_T B) \rhd (T + B))$. This last argument works, since an interpretation can be built in any arithmetical base theory: we could even have taken $T_0 = \mathsf{Q}$, where Q in Robinson's Arithmetic! Finally, composing (a) and (b), we get: $\Box_T((T + A) \rhd (T + B))$.

We can view P_0 as a 'P-ified' version of $\mathsf{J5}$. First note that we can rewrite $\mathsf{J5}$ as the equivalent:

$$- \vdash A \rhd \Diamond B \to A \rhd B$$

We get P_0, by putting a box in front of the consequent.

We can now pose a new conjecture: $\mathsf{ILW}^*\mathsf{P}_0 \stackrel{?}{=} \mathfrak{Il}(\mathsf{all})$.

4. Modal Semantics

4.1. Veltman Semantics

Frank Veltman, in response to questions of Albert Visser, found a purely modal Kripke style semantics for interpretability logic. Frank Veltman was working on conditionals at that time. However, $\rhd$ is not a proper conditional. See (Veltman, 1985).

Veltman's semantics extends the well-known Kripke semantics for L. Recall that an L-frame is a pair $\langle W, R\rangle$ where W is a nonempty set and R is a transitive conversely well-founded relation on W^2. An L-model is a triple $\langle W, R, \Vdash\rangle$ where $\langle W, R\rangle$ is an L-frame and $\Vdash$ is a forcing relation which commutes in the usual way with the connectives ($w \Vdash A \wedge B \Leftrightarrow w \Vdash A$ and $w \Vdash B$, etc.) and, moreover, $w \Vdash \Box A \Leftrightarrow \forall v\ (wRv \Rightarrow v \Vdash A)$. An IL-frame or Veltman frame is a triple $\langle W, R, \{S_w \mid w \in W\}\rangle$ such that:

1. $\langle W, R\rangle$ is an L-frame.
2. $S_w \subseteq w{\uparrow} \times w{\uparrow}$ ($w{\uparrow} := \{x \in W \mid wRx\}$).
3. $(R \restriction (w{\uparrow})) \subseteq S_w$.
4. S_w is reflexive.
5. S_w is transitive.

A Veltman model is a quadruple $\langle W, R, \{S_w \mid w \in W\}, \Vdash\rangle$. Here the triple $\langle W, R, \{S_w \mid w \in W\}\rangle$ is a Veltman frame and $\Vdash$ is a forcing relation with the extra condition that

– $w \Vdash A \rhd B \Leftrightarrow \forall u\ (w\,R\,u \Vdash A \Rightarrow \exists v\ \ u\,S_w\,v \Vdash B)$.
(We write e.g. '$u\,S_w\,v \Vdash B$' for 'uS_wv and $v \Vdash B$'.)

Veltman semantics is designed so that IL is sound and complete with respect to it.

4.2. Frames

Consider a frame $\mathcal{F} = \langle W, R, \{S_w \mid w \in W\}\rangle$. We define:

– $\mathcal{F} \models A :\Leftrightarrow$ for all forcing relations $\Vdash$, and for all $w{\in}W$, $w \Vdash A$.

– $\mathcal{F}$ is an ILW-frame if, for any x, $R; S_x$ is conversely well-founded. Here $u(R; S_x)v$ if, for some w, $uRwS_xv$.

– $\mathcal{F}$ is an an $\mathsf{ILM_0}$-frame if $xRyRzS_xuRv \Rightarrow yRv$.

– $\mathcal{F}$ is an $\mathsf{ILW^*}$-frame if it is both an ILW and an $\mathsf{ILM_0}$-frame.

– $\mathcal{F}$ is an ILM-frame if $yS_xzRu \Rightarrow yRu$.

– $\mathcal{F}$ is an ILP-frame if $xRyRzS_xu \Rightarrow zS_yu$.

We have the following correspondences: $\mathcal{F}$ is an ILW-frame, an $\mathsf{ILM_0}$-frame, an $\mathsf{ILW^*}$-frame, an ILM-frame, an ILP-frame if, respectively, $\mathcal{F} \models \mathsf{W}$, $\mathcal{F} \models \mathsf{M_0}$, $\mathcal{F} \models \mathsf{W^*}$, $\mathcal{F} \models \mathsf{M}$, $\mathcal{F} \models \mathsf{P}$.

4.3. Completeness Results

The logics IL, ILM, ILP are all modally complete with respect to their corresponding classes of frames. (See e.g. (Japaridze and de Jongh, 1998) or (Visser, 1998a).) In (de Jongh and Veltman, 20XX) it is shown that ILW is also modally complete. IL, ILW, ILM and ILP can be all shown to have the finite model property. It follows that they are decidable. In (Joosten, 1998) the modal completeness of $\mathsf{ILM_0}$ is proved. Although this theory is conjectured to be decidable too, its decidability is still open. Also, the question of the modal completeness for $\mathsf{ILW^*}$ remains open.

The arithmetical completeness of ILM and ILP was proved by embedding (the algebras associated with) the Veltman models for ILM, respectively ILP into the arithmetical theories. Thus the proofs of the arithmetical completeness theorem essentialy involved all three features: modal systems, Veltman semantics and arithmetical semantics.

4.4. The Story of $\mathsf{P_0}$

During fall of 1998, the progress in developing the modal completeness proof of $\mathsf{ILM_0}$ stagnated. It was thought that, perhaps, it would simplify things if we could strengthen the logic. Albert Visser tried to strengthen the frame condition of $\mathsf{ILM_0}$ to arrive at a stronger principle. Remember that the frame condition of $\mathsf{ILM_0}$ is:

- $xRyRzS_xuRv \to yRv$.

Instead of demanding an R-relation between x and v, one can demand an S_y-connection between z and v. If we have zS_yv, we must also have yRv, so indeed this move results in strengthening the frame condition. A corresponding principle, baptized $\mathsf{P_0}$, turns out to be:

$\mathsf{P_0} \vdash A \rhd \Diamond B \to \Box(A \rhd B)$.

Clearly every $\mathsf{ILP_0}$-frame is an $\mathsf{ILM_0}$-frame. If the logic $\mathsf{ILP_0}$ were modally complete then we would have: $\mathsf{ILP_0} \vdash M_0$ (i.e. $\mathsf{ILP_0}$ proves every instance of $\mathsf{M_0}$.) In (Joosten, 1998) it is shown that $\mathsf{ILP_0} \nvdash \mathsf{M_0}$ and hence that $\mathsf{ILP_0}$ is modally incomplete. The proof makes essential use of $\mathsf{IL_{set}}$-models which are a refinement of Veltman models, invented by Dick de Jongh. All logics are also sound w.r.t. the $\mathsf{IL_{set}}$-models, but more distinctions between principles become apparent. The main idea

is that S-relations don't run to a single world but to a set of worlds. More details can be found in (Joosten, 1998).

The real surprise was that the principle P_0 —which came from purely modal considerations— is valid in any reasonable arithmetical theory and hence should be in the core logic $\mathfrak{Il}(\mathsf{all})$.

4.5. New Principles by Modal Refinements

If we are looking for principles in $\mathfrak{Il}(\mathsf{all})$, we know for sure that they should be both in ILP and in ILM. A priori, there is an infinite search space but, Veltman models provide pretty good guidance in this quest. We shall make a convention on visualizing frame conditions. First, we do not represent all the relations in the pictures. If aRb and bRc are drawn, we will rather not draw the aRc that is dictated by transivity. So by a picture we actually mean its closure w.r.t. the closure conditions for Veltman models. Secondly, the R-relations will be drawn as straight lines and the S-relations as curved lines.

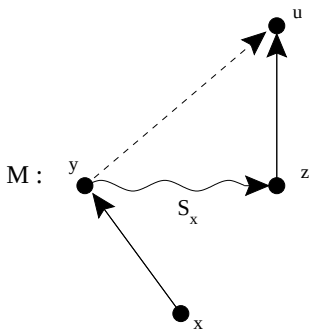


Figure 1.

If some specific relation is imposed by the principle whose frame condition we want to represent, we will indicate this by drawing a dashed line. Bearing this in mind we can visualize the frame condition of M. This condition was:

$$xRyS_xzRu \Rightarrow yRu.$$

In picture 1. this condition is represented. The imposed yRu is drawn as a dashed straight line.

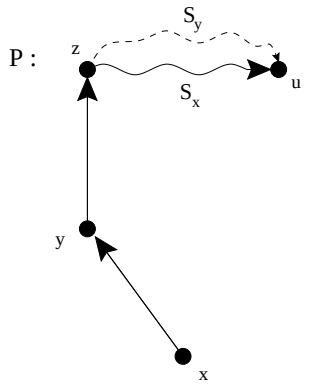


Figure 2.

When depicting the frame condition of P, we get a similar picture. The frame condition of P was:

$$xRyRzS_xu \Rightarrow zS_yu.$$

The imposed zS_yu is drawn as a curved intersected arrow. Note that the relations xRu and yRu are not shown although they have to exist.

A modal principle in $\mathfrak{Il}(\mathsf{all})$ should hold on all ILM- and ILP-frames. Consequently the frame condition of this principle should hold in both frame classes too. This was, of course, the case for all the principles considered so far. For example in both ILM-frames and ILP-frames we have that $R; S_x$ is conversely well-founded, for any x. And this was precisely the frame condition of W, a principle that holds in any reasonable arithmetical theory.

We can use the pictorial heuristic to guess new principles. The search space for new principles is thus confined to principles whose corresponding frame conditions are shared consequences of both the respective frame conditions of P and M. An example clarifies this concept.

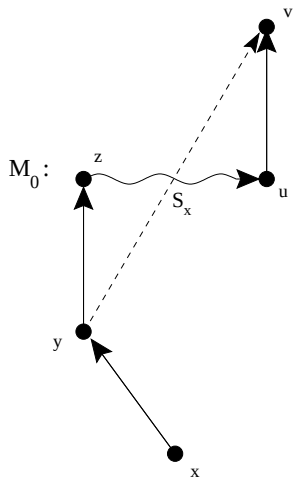


Figure 3.

A frame condition is drawn in picture 3. We assume $xRyRzS_xuRv$. Now the frame condition is that we impose yRv. The corresponding principle is readily found. It is $\mathsf{M}_0 \vdash A \rhd B \to (\Diamond A \wedge \Box C) \rhd (B \wedge \Box C)$. The R-relation between y and v is implied by the ILM frame condition because in an ILM frame we should have zRv and thus yRv as well. It is also implied by the frame condition of ILP because in an ILP-frame one has zS_yu and obviously also yRu. And this again yields yRv. The relation yRv is both in the closure of the frame under M and under P. So yRv is in the intersection.

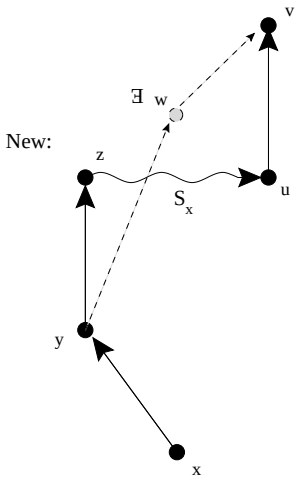


Figure 4.

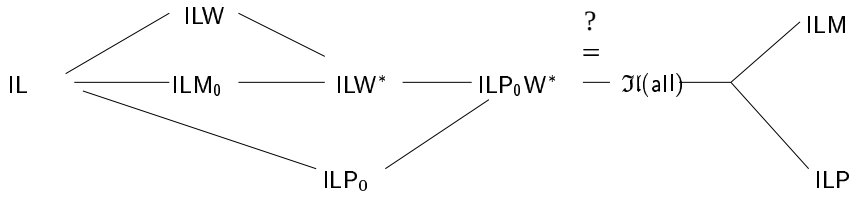
Reflection on the previous reasoning tells us that in both the ILM- as in the ILP-closure it is possible to go from y in two R-steps to v. In the M-case this is $yRzRv$ and in the P-case this is $yRuRv$. This idea could be captured by a somewhat different frame condition which demands the existence of an intermediate world w between y and v. This condition is represented in figure 4. So the frame condition is $xRyRzS_xuRv \Rightarrow \exists w\ yRwRv$ and a corresponding principle is $\vdash A \rhd B \rightarrow (\Diamond A \wedge \Box\Box C) \rhd (B \wedge \Box C)$. In this way we discover a new principle, say M_1, which is clearly stronger than M_0. *It is an open question whether* M_1 *is valid in all reasonable arithmetical theories!*

Another example of a principle in the intersection of ILM and ILP is the principle $A \rhd \Diamond B \Rightarrow \Box(A \rhd \Diamond B)$. In (Visser, 1998a) it is shown that this principle is *not* valid in all reasonable arithmetical theories.

5. Concluding Remarks

5.1. The Current Situation

At the moment of writing, we have a good picture of the relationships of the salient logics produced by our quest for $\mathfrak{Il}(\mathsf{all})$. There are the systems ILW and ILM_0 which are both modally complete and in the core logic we are looking for. Then there is the logic ILW^* which is the union of these two logics. It is conjectured to be decidable and complete but the problem is still open. The logic ILP_0 is completely independent from ILM_0, ILW and ILW^*. It is also in $\mathfrak{Il}(\mathsf{all})$. The union of all these logics, $\mathsf{ILP}_0\mathsf{W}^*$, is conjectured to be $\mathfrak{Il}(\mathsf{all})$.



5.2. Two Questions

We end our paper by formulating two questions of more restricted scope than our great problem.

Problem 1
The logic $\mathfrak{Il}(\mathsf{all})$ is in the intersection of ILM and ILP. But it cannot be equal to this system since e.g. the principle $A \rhd \Diamond B \to \Box(A \rhd \Diamond B)$ which is in the intersection, is not generally valid. A proof of this fact is given in (Visser, 1998a). The proof employs a heavy result due to Shavrukov, see (Shavrukov, 1997). Is there a more direct and more perspicuous proof of this fact?

Problem 2
Is the principle $A \rhd B \to (\Diamond A \wedge \Box\Box C) \rhd (B \wedge \Box C)$ arithmetically valid? It is certainly in ILM and ILP and thus valid both in essentially reflexive and in finitely axiomatized reasonable arithmetical theories. Moreover it can be shown to be valid for $I\Delta_0 + B\Sigma_1 + \Omega_1$. Yet it is hard to see why it should be generally valid. In fact we conjecture that it is not.

Acknowledgements

We thank Dick de Jongh for many enlightening conversations. We thank Rosalie Iemhoff for her careful reading of the penultimate draft.

Notes

[1] More generally, we can use $\delta(\vec{x})$, using several variables to represent one object. We are mainly interested in theories with sequence coding in which we can restrict ourselves to δ with just one free variable.

[2] To make this work in our set-up, we have to assume a version of Euclidean geometry with a constant for a point and an axiom stating that the point is indeed a point. To avoid the necessity of such inelegant stipulations we have to improve a bit on our present definition of interpretation.

[3] Lev Beklemishev places this kind of study between structural prooftheory which studies specific proof systems and proofs, and recursion theoretic prooftheory where theories are considered as RE sets of theorems. Here we abstract away from many details of the proof system and from detailed proofs, however e.g. the defining formula of the set of axioms is a feature that can make a difference. Perhaps one could say that the modal study of provability and interpretability is part of *intensional prooftheory*.

[4] For the modal language restricted to the *unary* connective $\top \rhd A$ in combination with $\Box$, the problem of the interpretability logic of all theories has been solved by Maarten de Rijke, see his (de Rijke, 1992).

[5] We can also interpret $\rhd$ as partial conservativity w.r.t. a suitable class of formulas.

[6] We can derive $\mathsf{L4}$ without using $\mathsf{L3}$ by employing a surprising argument of Kreisel (presented in (Smoryński, 1977)). However, this argument includes the verification of $\mathsf{J5}$.

[7] André van Kooy showed in his masters thesis (Department of Philosophy, Utrecht University) that for finitely axiomatized theories in a relational language one can make the transformation of a proof of A into a proof of the provability of A *linear time*. This seems to be only possible via the J4,5-route.

[8] We are not quite sure who found this fact first. It might be well have been discovered independently by Friedman, Pudlák and Solovay.

[9] In fact the class is somewhat bigger. The reader is referred to (Visser, 1998a) for further elaboration.

References

Areces, C., D. de Jongh, and E. Hoogland: 1998, 'The Interpolation Theorem for IL and ILP'. In: *Proceedings of AiML98. Advances in Modal Logic*. Uppsala. Sweden.

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

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

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

Buss, S.: 1986, *Bounded Arithmetic*. Bibliopolis, Napoli.

de Jongh, D. and F. Veltman: 1990, 'Provability logics for relative interpretability'. In: (Petkov, 1990). pp. 31–42.

de Jongh, D. and F. Veltman: 20XX, 'Modal Completeness of *ILW*'. Unpublished.

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

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

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

Iemhoff, R.: 20XX, 'A Modal Analysis of Some Principles of the Provability Logic of Heyting Arithmetic'. In: *Proceedings of AiML'98*, Vol. 2. Uppsala.

Japaridze, G. and D. de Jongh: 1998, 'The logic of provability'. In: S. Buss (ed.): *Handbook of proof theory*. North-Holland Publishing Co., amsterdam edition, pp. 475–546.

Joosten, J.: 1998, *Towards the Interpretability Logic of all Reasonable Arithmetical Theories*. Master's Thesis, ILLC, University of Amsterdam.

Lindström, P.: 1997, *Aspects of Incompleteness*, Vol. Lecture Notes in Logic 10. Berlin: Springer.

Petkov, P. (ed.): 1990, 'Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria'. Plenum Press, Boston.

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

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

Shavrukov, V.: 1997, 'Interpreting Reflexive Theories in Finitely Many Axioms'. *Fundamenta Mathematicae* **152**, 99–116.

Smoryński, C.: 1977, 'The Incompleteness Theorems'. In: J. Barwise (ed.): *Handbook of Mathematical Logic*. Amsterdam: North-Holland, pp. 821–865.

Solovay, R.: 1976, 'On interpretability in set theories'. Unpublished manuscript.
Švejdar, V.: 1978, 'Degrees of interpretability'. *Commentationes Mathematicae Universitatis Carolinae* **19**, 789–813.
Tarski, A., A. Mostowski, and R. Robinson: 1953, *Undecidable theories.* North–Holland, Amsterdam.
Veltman, F.: 1985, 'Logic for conditionals'. Ph.D. thesis, Department of Philosophy, University of Amsterdam.
Verbrugge, L.: 1993, *Efficient metamathematics.* ILLC-disseration series 1993-3, Amsterdam.
Visser, A.: 1985, 'Evaluation, provably deductive equivalence in Heyting's Arithmetic of substitution instances of propositional formulas'. Technical Report LGPS 4, Department of Philosophy, Utrecht University.
Visser, A.: 1990, 'Interpretability logic'. In: *(*Petkov, 1990*)*. pp. 175–209.
Visser, A.: 1991, 'The formalization of interpretability'. *Studia Logica* **51**, 81–105.
Visser, A.: 1994, *Propositional combinations of Σ-sentences in Heyting's Arithmetic*, Logic Group Preprint Series 117. Utrecht: Department of Philosophy, Utrecht University.
Visser, A.: 1998a, 'An Overview of Interpretability Logic'. In: M. Kracht, M. de Rijke, H. Wansing, and M. Zakharyaschev (eds.): *Advances in Modal Logic, vol 1*, CSLI Lecture Notes, no. 87. Stanford: Center for the Study of Language and Information, pp. 307–359.
Visser, A.: 1998b, 'Rules and Arithmetics'. Logic Group Preprint Series 186, Department of Philosophy, Utrecht University, Heidelberglaan 8, 3584 CS Utrecht.
Visser, A., J. van Benthem, D. de Jongh, and G. R. de Lavalette: 1995, 'NNIL, a Study in Intuitionistic Propositional Logic'. In: A. Ponse, M. de Rijke, and Y. Venema (eds.): *Modal Logic and Process Algebra, a Bisimulation Perspective*, CSLI Lecture Notes, no. 53. Stanford: Center for the Study of Language and Information, pp. 289–326.
Wilkie, A. and J. Paris: 1987, 'On the scheme of of induction for bounded arithmetic formulas'. *Annals of Pure and Applied Logic* **35**, 261–302.
Zambella, D.: 1992, 'On the proofs of arithmetical completeness of interpretability logic'. *The Notre Dame Journal of Formal Logic* **35**, 542–551.

Address for Offprints: Department of Philosophy
Utrecht University
Heidelberglaan 8
3584 CS Utrecht