Notre Dame Journal of Formal Logic
Volume 47, Number 2, 2006

Interpolation and the Interpretability Logic of PA

Evan Goris

Abstract In this paper we will be concerned with the interpretability logic of PA and in particular with the fact that this logic, which is denoted by ILM, does not have the interpolation property. An example for this fact seems to emerge from the fact that ILM cannot express Σ_1-ness. This suggests a way to extend the expressive power of interpretability logic, namely, by an additional operator for Σ_1-ness, which might give us a logic with the interpolation property. We will formulate this extension, give an axiomatization which is modally complete and arithmetically complete (although for proofs of these theorems we refer to an earlier paper), and investigate interpolation. We show that this logic still does not have the interpolation property.

1 Introduction

In this paper[1] we will be concerned with what are known as interpretability logics (Visser [16]). These are nonstandard extensions of the well-known modal logic GL. As is well known, we can interpret modal formulas by reading the propositional variables as arbitrary arithmetical sentences and the $\Box$ as T-provable for some arithmetical theory T. GL turns out to be a complete axiomatization for this semantics (see Solovay [15] and Boolos [4]). This idea can be naturally extended by introducing new modal operators and giving them an arithmetical meaning. One such extension is interpretability logic. The language of this logic contains, besides the $\Box$, a binary modal operator $\rhd$ (see [16], Visser [17], and Japaridze and de Jongh [12]) (we introduce the arithmetical meaning of this operator below). Although GL enjoys the interpolation property [4], this feature, although it does occur sometimes (see Areces et al. [1]), is very special for interpretability logics (indeed, another difference is that GL seems to be rather independent of the theory T, whereas we obtain different interpretability logics when we vary T).

Received March 12, 2005; accepted October 3, 2005; printed July 20, 2006
2000 Mathematics Subject Classification: Primary, 03B45; Secondary, 03F30
Keywords: provability logic, interpretability logic, interpolation

As interpolation can be seen as to how well behaved a proof system is, and the modal study of interpretability logics is already rather complex (de Jongh and Veltman [5], Joosten and Goris [13]), it might be of interest to determine extensions that do have interpolation.[2] We will examine this question in this paper for the case $T = \mathrm{PA}$. The interpretability logic of PA, which is defined below, is denoted by ILM.

As we will see, a counterexample due to Ignatiev for interpolation in ILM suggests that extending the language of interpretability logic with an additional operator that expresses Σ_1-ness might give us a way to formulate a logic that can talk about interpretability and has the interpolation property. This "suggestion" is exactly what we will investigate in this paper: we formulate such an extension, give a modally complete and arithmetically complete axiomatization, and investigate interpolation. However, and this is the main result of this paper, we will show that we still do not have the interpolation property.

As for the failure of interpolation of this extension the paper is self contained. Proofs of the two completeness theorems (arithmetical and modal) can be found in Goris [9].

The organization of this paper is as follows. In Section 2 we introduce interpretability logic and mention the main results and agree on some notations in Section 3. In Section 4 we define the arithmetical reading of interpretability logic. In Section 5 we give some more detailed motivation for our study and present an example for failure of interpolation in ILM. In Section 6 we formulate the extension of interpretability logic with an additional operator for Σ_1-ness. We give a modally complete and arithmetically complete axiomatization, which we call ILM(S),[3] prove that this logic does not have the interpolation property, and indicate the gap in the expressiveness of ILM(S) that seems to be the reason for this.

2 Preliminaries: Interpretability Logics

Definition 2.1 (IL-formulas) IL-*formulas* are built up using some fixed set of propositional variables, the propositional connectives, a unary modal operator $\Box$, and a binary modal operator $\rhd$.

With regard to priorities, $\rhd$ behaves similarly as $\rightarrow$, although $\rhd$ binds stronger. So $A \wedge B \rhd C$ means $(A \wedge B) \rhd C$ and $A \rightarrow B \rhd C$ means $A \rightarrow (B \rhd C)$. As usual, we use $\Diamond$ as an abbreviation for $\neg\Box\neg$. (As we will see later, we can use $\Box A$ as an abbreviation for $\neg A \rhd \bot$.)

Let us sketch how we can extend the arithmetical meaning of standard modal formulas to IL-formulas (a precise definition will be presented below). If A and B are IL-formulas and A^* and B^* arithmetical sentences, the "arithmetical meaning" of A and B, respectively, then the arithmetical meaning of $A \rhd B$ is a formalization of

$$\mathrm{PA} + A^* \text{ interprets } \mathrm{PA} + B^*.$$

In general, a theory T interprets a theory S if there exists a translation of formulas of S into formulas of T such that T proves all the (translations of the) theorems of S. For a precise formulation see [12] or [17]. We will not bother with this here since we will switch to another, but equivalent over PA, arithmetical reading of $\rhd$ anyway.

Definition 2.2 (IL) With IL we will refer to the following set of axiom schemata.

L1 $\quad \Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$,

L2 $\Box A \to \Box\Box A$,
L3 $\Box(\Box A \to A) \to \Box A$,
J1 $\Box(A \to B) \to A \rhd B$,
J2 $(A \rhd B) \wedge (B \rhd C) \to A \rhd C$,
J3 $(A \rhd C) \wedge (B \rhd C) \to A \vee B \rhd C$,
J4 $A \rhd B \to (\Diamond A \to \Diamond B)$,
J5 $\Diamond A \rhd A$.

We obtain *the logic* IL by taking as axioms all instances of the above schemata, the classical propositional tautologies in the enriched language, and close off under necessitation and modus ponens. We write IL $\vdash A$ for $A \in$ 'the logic IL'. Without danger of confusion we speak of IL when we mean the logic IL.

Before we move on, let us state a well-known but important lemma (for a proof see [12]).

Lemma 2.3

1. IL $\vdash A \rhd A \wedge \Box\neg A$.
2. IL $\vdash \Diamond A \vee A \rhd A$.
3. IL $\vdash \Box A \leftrightarrow \neg A \rhd \bot$.

Note that by Item 3 above, when working with extensions of IL, we can treat $\Box$ as a defined symbol.

If T is an arithmetical theory then, as is the case in provability logic, we say that an IL-formula A is *T-valid* when, independent of the choice of the arithmetical meaning of the propositional variables, A translates to a T-provable formula. When T is understood, these formulas are also referred to as *always provable formulas* or as *valid formulas*. All this will be made more precise in Section 4 below.

For certain theories T, the class of T-valid IL-formulas can be axiomatized by adjoining to IL appropriate axiom schemata. PA is a theory for which such a schema has been obtained.

Definition 2.4 (ILM, M) With M we denote the schema,

$$A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C.$$

ILM is the set of schemata IL + {M} and we obtain *the logic* ILM by taking as axioms all instances of ILM, the classical propositional tautologies, and close off under necessitation and modus ponens. Again we write ILM when we mean the logic ILM.

We can evaluate IL-formulas on Veltman frames and in Veltman models.

Definition 2.5 (Veltman Frame) A *Veltman frame*, or simply a *frame*, is a triple $F = \langle W, R, S\rangle$ where

1. $\langle W, R\rangle$ is a GL-frame (in other words, W is a set and R is a transitive, conversely well-founded binary relation on W);
2. S is a ternary relation on W; with S_w we designate the binary relation $\{(a, b) \mid (w, a, b) \in S\}$; additionally, we require for all a, b, c, w, t that the following holds:
   (a) $aS_wb \Rightarrow wRa \;\&\; wRb$,
   (b) $wRaRb \Rightarrow aS_wb$,
   (c) $wRa \Rightarrow aS_wa$.

Remark Usually one postulates in addition that each S_w is transitive. For technical reasons we translated this property into the forcing relation defined below. We cannot say much about this at this point; see the paragraphs at the end of Subsection 6.1.

Definition 2.6 (Veltman model) A *Veltman model*, or simply a *model*, is a quadruple $M = \langle W, R, S, \Vdash \rangle$ where $\langle W, R, S \rangle$ is a Veltman frame and $\Vdash$ is a *forcing relation* between elements of W and IL-formulas that satisfies the following requirements.

1. $w \Vdash A \rhd B$ iff for each wRu such that $u \Vdash A$, there exists a v such that $u(S_w)^* v$ and $v \Vdash B$ (here $(S_w)^*$ is the reflexive, transitive closure of S_w).
2. $\Vdash$ commutes with Boolean connectives, for example, $w \Vdash A \wedge B$ iff $w \Vdash A$ and $w \Vdash B$.

Notice that by the equivalence $\Box A \leftrightarrow \neg A \rhd \bot$ we obtain the usual forcing condition for $\Box$.

We say that an IL-formula A is *valid on a frame* $F = \langle W, R, S \rangle$, and write $F \models A$, whenever for any Veltman model $M = \langle W, R, S, \Vdash \rangle$ (we say that such a model is *based on* F) and any $m \in M$ we have $m \Vdash A$.

Definition 2.7 (ILM-frame, ILM-model) A frame that additionally satisfies the following property is called an ILM-frame.

$$\forall waa'b\ (a(S_w)^* a' Rb \Rightarrow aRb).$$

An ILM-model is a Veltman model that is based on a ILM-frame.

Theorem 2.8 (Modal completeness of IL and ILM [6], [12]) *For any modal formula A we have the following.*

1. *A is valid on all frames iff* $\mathrm{IL} \vdash A$.
2. *A is valid on all* ILM*-frames iff* $\mathrm{ILM} \vdash A$.

3 Notations

In this section we agree on some notations and conventions. Uppercase characters $A, B, C, \ldots$ range over modal formulas. The lowercase characters $a, b, c, \ldots$, $p, q, r, \ldots$ denote propositional variables and nodes in frames and models (no confusion will arise).

For models M we will use the notation M for both the model and its domain, similarly for frames. If $F = \langle W, R, S \rangle$ then we write W^F for W, R^F for R, S^F for S.

If A is a modal formula then $\boxdot A =_{\text{def}} \Box A \wedge A$. When we want to apply an operator, say $\boxdot$, to all formulas in some finite set Γ then we write $\boxdot\Gamma$.

Finally one should note that some objects occurring in this paper carry slightly different names than they did in [9]. That is, ILM(S) was Σ_1ILM, SL was Σ_1L, ILS-formula was Σ_1ILM-formula, and M(S) was M(Σ_1).

4 Arithmetical Reading of Modal Formulas

All the modal logics presented in this paper have an arithmetical reading. Although we are mainly concerned with interpolation in the modal logics, many concepts are motivated out of the arithmetical semantics of these logics. Therefore we elaborate a bit on the arithmetical side in this section.

Lowercase Greek letters like φ and ψ denote first-order formulas with identity in the language of PA: $\langle +, \times, 0, 1\rangle$. As usual, we take $\rightarrow$ and $\forall$ as logical symbols, fix $\bot$ to be some provably false sentence (like $0 = 1$), and treat the other symbols as defined. Boldface characters like $\mathbf{n}$ and $\mathbf{w}$ denote fixed (standard) natural numbers. *Numerals* are canonical representations of standard natural numbers in the language of PA. If $\mathbf{n}$ is a natural number, then with $\underline{\mathbf{n}}$ we denote its corresponding numeral and it is recursively defined as follows: $\underline{\mathbf{0}} = 0$, $\underline{\mathbf{n+1}} = \underline{\mathbf{n}} + 1$. Normal characters like n and w are (just) variables.

We assume a standard coding of the syntax of PA in PA (see, for example, [4]). If x is a syntactic object, then we denote by $\ulcorner x \urcorner$ its code. For the different syntactic objects we have Δ_1-formulas that define the codes of those objects in PA. For example, we have a Δ_1-formula $\mathrm{Formula}(x)$, which is provable of $\mathbf{n}$ if and only if $\mathbf{n} = \ulcorner \varphi \urcorner$ for some formula φ. With $\dot{\neg}$ we denote a primitive recursive function such that for each formula φ, $\dot{\neg}\ulcorner \varphi \urcorner = \ulcorner \neg \varphi \urcorner$. Similar conventions hold for the other Boolean connectives. It is well known that we can set things up such that PA can prove the recursive properties of syntactic objects. For example, PA proves that when x is a formula, then so is $\dot{\neg} x$.

We let $\Box(x)$ be a standard Gödel provability predicate [4]. So, in particular, $\Box(x)$ is a Σ_1-formula, provably false of any $\mathbf{n}$ which is not the code of some formula and for which we have

$$\mathrm{PA} \vdash \varphi \Leftrightarrow \mathrm{PA} \vdash \Box(\underline{\ulcorner \varphi \urcorner}). \tag{1}$$

In what follows, for readability, we write $\Box\ulcorner \varphi \urcorner$ for $\Box(\underline{\ulcorner \varphi \urcorner})$, or even more generally, we will identify standard natural numbers with their corresponding numerals. Besides (1), the following conditions (Löb derivability conditions [4]) are well known. For all formulas φ, ψ,

1. $\mathrm{PA} \vdash \Box\ulcorner \varphi \urcorner \rightarrow \Box\ulcorner \Box\ulcorner \varphi \urcorner \urcorner$,
2. $\mathrm{PA} \vdash \Box\ulcorner (\varphi \rightarrow \psi) \urcorner \rightarrow (\Box\ulcorner \varphi \urcorner \rightarrow \Box\ulcorner \psi \urcorner)$,
3. $\mathrm{PA} \vdash \Box\ulcorner \Box\ulcorner \varphi \urcorner \rightarrow \varphi \urcorner \rightarrow \Box\ulcorner \varphi \urcorner$.

When working with formalized provability, generalized versions of the above conditions are often necessary. For us the following two observations about $\Box(x)$ are sufficient. For one we have the following generalization of Item 2.

$$\mathrm{PA} \vdash \Box(x \dot{\rightarrow} y) \rightarrow (\Box(x) \rightarrow \Box(y)).$$

And additionally we have that if $t(x_1, \ldots, x_k)$ is a term in the language $\{\dot{\rightarrow}, \ulcorner \bot \urcorner\}$ and for all $\varphi_1, \ldots, \varphi_k$, $t(\ulcorner \varphi_1 \urcorner, \ldots, \ulcorner \varphi_k \urcorner)$ is the code of some propositional tautology, then

$$\mathrm{PA} \vdash \bigwedge_{1 \leq i \leq k} \mathrm{Formula}(x_i) \rightarrow \Box(t(\overline{x})).$$

Stated informally, the above says that we can provably do any propositional reasoning under the $\Box$.

Using this standard provability predicate $\Box(x)$ we can translate IL-formulas that only contain the $\Box$ as a modal operator to arithmetical sentences. To give an arbitrary IL-formula arithmetical meaning we will use a formalization of the following metamathematical concept.

Definition 4.1 (Π_1**-conservativity**) Let S and T be finite extensions of PA. We say that S is Π_1-conservative over T if for any Π_1-sentence π we have that $S \vdash \pi$ implies $T \vdash \pi$.

In order to formalize this notion we define

$$\Sigma_1(x) =_{\text{def}} \exists y(\Sigma_1!(y) \wedge \Box(x \dot{\leftrightarrow} y)).$$

Here $\Sigma_1!(x)$ is a Δ_1-formula that defines the codes of strict Σ_1-formulas. And we define a binary predicate $\rhd$, formalizing Π_1-conservativity, as follows:

$$x \rhd y =_{\text{def}} \forall z(\Pi_1(z) \to (\Box(y \dot{\to} z) \to \Box(x \dot{\to} z))),$$

where $\Pi_1(x)$ is a shorthand for $\Sigma_1(\dot{\neg} x)$.

Now we can give arithmetical meaning to IL-formulas as follows. An *arithmetical translation* is a function $*$ from modal formulas A to arithmetical sentences A^* that satisfies the following.

1. $\bot^* = \bot$,
2. $(A \to B)^* = A^* \to B^*$,
3. $(A \rhd B)^* = \ulcorner A^* \urcorner \rhd \ulcorner B^* \urcorner$.

Notice that $*$ is uniquely determined when we know p^* for all propositional variables p and that $(\Box A)^* (= (\neg A \rhd \bot)^*)$ is equivalent to $\Box(\ulcorner A^* \urcorner)$. An ILS-formula is *arithmetically valid*, or simply *valid*, when $\text{PA} \vdash A^*$ for any arithmetical translation $*$. The following theorem is the main motivation behind ILM.

Theorem 4.2 (Arithmetical completeness of ILM [3], [14][4]) $\text{ILM} \vdash A$ *if and only if for all* $*$, $\text{PA} \vdash A^*$.

5 Motivation

In this section we analyze interpolation in ILM. As a preparation let us see why (each instantiation of) the M schema is true. That is, we show that for all first-order formulas φ, ψ, and η,

$$\mathbf{N} \models \ulcorner \varphi \urcorner \rhd \ulcorner \psi \urcorner \to \ulcorner \varphi \wedge \Box \ulcorner \eta \urcorner \urcorner \rhd \ulcorner \psi \wedge \Box \ulcorner \eta \urcorner \urcorner. \tag{2}$$

Suppose $\ulcorner \varphi \urcorner \rhd \ulcorner \psi \urcorner$. Also suppose that π is some Π_1-sentence provable in $\text{PA} + \psi \wedge \Box \ulcorner \eta \urcorner$. Then $\text{PA} + \psi$ proves the Π_1-sentence $\Box \ulcorner \eta \urcorner \to \pi$, and thus $\text{PA} + \varphi$ proves $\Box \ulcorner \eta \urcorner \to \pi$ as well; conclusion: $\text{PA} + \varphi \wedge \Box \ulcorner \eta \urcorner$ proves π.

Now let us consider a well-known counterexample, due to Ignatiev, for interpolation in ILM. We have

$$\text{ILM} \vdash \Box(p \leftrightarrow \Box q) \to (r \rhd s \to r \wedge p \rhd s \wedge p), \tag{3}$$

but for any I, with propositional variables among $\{p\}$, we do not have both

$$\text{ILM} \vdash \Box(p \leftrightarrow \Box q) \to I$$

and

$$\text{ILM} \vdash I \to (r \rhd s \to r \wedge p \rhd s \wedge p).$$

For a proof see [17].[5]

Suppose we could express Σ_1-ness by an IL-formula with one propositional variable, say $\Sigma_1(p)$. The above proof that the M schema is true actually shows that the following schema is true:

$$\Sigma_1(C) \to (A \rhd B \to A \wedge C \rhd B \wedge C).$$

Moreover, that argument can be carried out in PA and that shows that these principles are arithmetically valid. Since $\Box(C \leftrightarrow \Box D) \to \Sigma_1(C)$ is arithmetically valid as

well, and ILM is known to prove all arithmetically valid formulas, we would have an interpolant for (3) (namely, $\Sigma_1(p)$).

The rest of this paper is devoted to adjoining an operator to ILM, for which we stipulate that its arithmetical meaning is Σ_1-ness.

6 The Logic ILM(S)

In this section we develop a modal logic which can talk about Π_1-conservativity and Σ_1-ness. We give a modally complete and arithmetically complete axiomatization and give a counterexample for interpolation.

Definition 6.1 (ILS-formulas, SL-formulas) ILS-*formulas* are built up using some fixed set of propositional variables, the propositional connectives, unary modal operators $\Box$ and Σ_1, and the binary operator $\rhd$. ILS-formulas that do not contain $\rhd$ will be referred to as SL-*formulas*.

Definition 6.2 (SL) With SL we denote the following set of schemata.

L1 $\Box(A \to B) \to (\Box A \to \Box B)$,
L2 $\Box A \to \Box\Box A$,
L3 $\Box(\Box A \to A) \to \Box A$,
S1 $\Sigma_1 A \wedge \Sigma_1 B \to \Sigma_1(A \wedge B)$,
S2 $\Sigma_1 A \wedge \Sigma_1 B \to \Sigma_1(A \vee B)$,
S3 $\Sigma_1 A \wedge \Box(A \leftrightarrow B) \to \Sigma_1 B$,
S4 $\Sigma_1 A \to \Box\Sigma_1 A$,
S5 $\Sigma_1 \bot$,
S6 $\Sigma_1 \Box A$,
S7 $\Sigma_1 \Sigma_1 A$,
S8 $\Sigma_1 A \to \Box(A \to \Box A)$.

By translating SL-formulas in the obvious way to arithmetical sentences (that is, by translating the new modal operator Σ_1 to a formula $\Sigma_1(x)$, which asserts that x is the code of a Σ_1-formula) it is shown in [9] that SL is arithmetically complete. Also let us note that SL is more expressive than GL. This follows immediately from the fact that GL has the interpolation property whereas SL does not (this is shown in [9] but also follows from Section 6.1 below). So in particular the implication in S8 does not reverse. (This latter fact already follows from a result by Beklemishev [2].) However, in this paper will we skip an explicit treatment of SL and immediately move on with the full range of ILS-formulas.

Definition 6.3 (ILM(S)) ILM(S) is the set of schemata IL + SL together with the schema M(S):

$$\Sigma_1 C \wedge (A \rhd B) \to A \wedge C \rhd B \wedge C.$$

The logic ILM(S) is the smallest set of ILM(S)-formulas which contains all instances of all the ILM(S) schemata and is closed under modus ponens and necessitation. We write ILM(S) $\vdash A$, or $\vdash A$, for $A \in$ *the logic* ILM(S).

We will evaluate ILS-formulas on Veltman frames and Veltman models. The appropriate Veltman frames for ILM(S) turns out to be the class of ILM-frames.

Definition 6.4 (ILM(S)-frame, ILM(S)-model) An ILM(S)-*frame* is an ILM-frame. An ILM(S)-model is a quadruple $\langle W, R, S, \Vdash\rangle$ such that $\langle W, R, S\rangle$ is an

ILM(S)-frame and $\Vdash$ is a relation between elements of W and ILS-formulas that satisfies the following requirements.

1. For the Boolean connectives and $\rhd$ the same clauses as for Veltman models apply.
2. $w \Vdash \Sigma_1 A$ iff for all w' such that $w(R \cup \bigcup_{x \in M} S_x)^* w'$ and all v and u such that $v S_{w'} u$, we have $v \Vdash A \Rightarrow u \Vdash A$.

Of course we can talk in this setting of frame validity as well. And we have the following theorem.

Theorem 6.5 (Modal completeness) ILM(S) $\vdash A$ *if and only if* $F \models A$ *for any* ILM(S)*-frame* F.

Proof We will show the soundness direction since we will need it later. For the completeness direction see [9].

As usual it suffices to show that each instance of an axiom schema is valid on each ILM(S) frame. So let $M = \langle W, R, S, \Vdash \rangle$ be some model based on an ILM(S)-frame. Since $\langle W, R \rangle$ is a GL-frame, axioms L1, L2, and L3 are known to hold [4].

S1 and S2: Let $w \in W$ and suppose $w \Vdash \Sigma_1 A$ and $w \Vdash \Sigma_1 B$. Let $w', x, y \in W$ with $w(R \cup \bigcup_{u \in W} S_u)^* w'$ and $x S_{w'} y$. In case $x \Vdash A \wedge B$ then both $x \Vdash A$ and $x \Vdash B$ and thus $y \Vdash A \wedge B$. In case $x \Vdash A \vee B$ then $x \Vdash A$ or $x \Vdash B$. In the former case $y \Vdash A$ and thus $y \Vdash A \vee B$. In the latter $y \Vdash B$ and thus $y \Vdash A \vee B$.

S3: Let $w, w', x, y \in W$ and suppose $w \Vdash \Sigma_1 A$, $w \Vdash \Box(A \leftrightarrow B)$, $w(R \cup \bigcup_{u \in W} S_u)^* w'$, $x S_{w'} y$, and $x \Vdash B$. From $x S_{w'} y$ we get $w' R x, y$ and thus by the M-property for M and/or transitivity of R we get $w R x, y$. So we get $x \Vdash A$, which gives $y \Vdash A$ and thus $y \Vdash B$.

S4: Let $w \in W$ and suppose $w \Vdash \Sigma_1 A$. Let $v \in W$ with $w R v$. We have to show that $v \Vdash \Sigma_1 A$. So take $v', x, y \in W$ with $v(R \cup \bigcup_{u \in W} S_u)^* v'$, $x S_{v'} y$, and $x \Vdash A$. We clearly have $w(R \cup \bigcup_{u \in M} S_u)^* v'$. And thus, indeed, $y \Vdash A$.

S5: This is clear.

S6: Let $w, w', x, y \in W$ with $w(R \cup \bigcup_{u \in W} S_u)^* w'$, $x S_{w'} y$, and $x \Vdash \Box A$. We have to show that $y \Vdash \Box A$. So pick $z \in W$ with $y R z$. By the M-frame property we have $x R z$ and thus $z \Vdash A$.

S7: Let $w, x, y \in W$ and suppose $x S_w y$, $x \Vdash \Sigma_1 A$. We will show that $y \Vdash \Sigma_1 A$. So let y' be such that $y(R \cup \bigcup_{u \in W} S_u)^* y'$. We then also have $x(R \cup \bigcup_{u \in W} S_u)^* y'$ and thus if $z S_{y'} z'$ and $z \Vdash A$ we obtain from $x \Vdash \Sigma_1 A$ that $z' \Vdash A$.

S8: Let $w \in W$ and suppose $w \Vdash \Sigma_1 A$. Let $x \in W$ with $w R x \Vdash A$. We have to show that $x \Vdash \Box A$. So let $y \in W$ with $x R y$. By Item 2b of the definition of a Veltman frame we have $x S_w y$ and thus by $w \Vdash \Sigma_1 A$ we get $y \Vdash A$.

J1: Let $w \in W$ such that $w \Vdash \Box(A \to B)$. Take $x \in W$ such that $w R x$ and $x \Vdash A$. We have to show that for some y with $x(S_w)^* y$ we have $y \Vdash B$. Since $x \Vdash A \to B$ we can take $y = x$.

J2: Let $w \in W$ such that $w \Vdash (A \rhd B) \wedge (B \rhd C)$. Take $x \in W$ such that $w R x$ and $x \Vdash A$. We have to show that for some y with $x(S_w)^* y$ we have $y \Vdash C$. Since $w \Vdash A \rhd B$ we find some y' with $x(S_w)^* y'$ and $y' \Vdash B$. Thus since $w \Vdash B \rhd C$ there exists some y with $y'(S_w)^* y$, and thus also $x(S_w)^* y$, with $y \Vdash C$.

J3: Let $w \in W$ and suppose $w \Vdash (A \rhd C) \wedge (B \rhd C)$. Let $x \in W$ such that wRx and $x \Vdash A \vee B$. We need a y with $x(S_w)^*y$ and $y \Vdash C$. In case $x \Vdash A$ we find a required y using $w \Vdash A \rhd C$ and in case $x \Vdash B$ we use $w \Vdash B \rhd C$.

J4: Let $w \in W$ and suppose $w \Vdash A \rhd B$ and $w \Vdash \Diamond A$. Thus for some x with wRx we have $x \Vdash A$ and thus there exists a y with $x(S_w)^*y$ and $y \Vdash B$. By Item 2a of the definition of a Veltman frame we get wRy and thus $w \Vdash \Diamond B$.

J5: Let $w, x \in W$ with wRx and suppose $x \Vdash \Diamond A$. Thus for some y with xRy we have $y \Vdash A$. By Item 2b of the definition of a Veltman frame we have xS_wy and thus we have shown $w \Vdash \Diamond A \rhd A$.

M(S): Suppose $w \Vdash \Sigma_1 C$ and $w \Vdash A \rhd B$. If wRx, $x \Vdash A \wedge C$ then for some y with $x(S_w)^*y$ we have $y \Vdash B$. Since $w \Vdash \Sigma_1 C$ implies that C is "preserved along S_w" we get $y \Vdash C$. □

An obvious adaption of the definition of an arithmetical translation for IL-formulas to ILS-formulas gives us the following theorem. A proof can be found in [9] and is basically a combination of the arithmetical completeness proof for ILM ([14] and [3]; see also [12]) and the one for HGL [8].

Theorem 6.6 (Arithmetical completeness) ILM(S) $\vdash A$ *if and only if for all arithmetical translations* $*$ *we have* PA $\vdash A^*$.

6.1 Interpolation

Theorem 6.7 (Failure of interpolation) *There exist formulas A and B such that* ILM(S) $\vdash A \to \neg B$ *but for no formula I which contains only propositional variables contained in both A and B we have* ILM(S) $\vdash A \to I$ *and* ILM(S) $\vdash I \to \neg B$.

Consider the following two formulas.

$$A_s = A(p,q,s) =_{\text{def}} \neg\Sigma_1 q \wedge \Sigma_1 s \wedge \Box(s \to q) \wedge \Box(p \wedge q \to s),$$
$$B_r = B(p,q,r) =_{\text{def}} \neg\Sigma_1 q \wedge \Sigma_1 r \wedge \Box(r \to q) \wedge \Box(\neg p \wedge q \to r).$$

Before we prove that these formulas constitute a counterexample for the interpolation property, let us see what would be needed for an interpolant to exist. A sufficient addition for an interpolant is Σ_1-interpolability (Σ_1-interpolability was first investigated in [11]), which is formalized by the following first-order formula.

$$I_{\Sigma_1}(x,y) =_{\text{def}} \exists z(\Sigma_1(z) \wedge \Box(x \dot{\to} z) \wedge \Box(z \dot{\to} y)).$$

Now let $*$ be any arithmetical translation. Let us first make three obvious observations and prove a lemma.

$$\text{PA} \vdash A(p,q,s)^* \to I_{\Sigma_1}(\ulcorner p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner). \tag{4}$$
$$\text{PA} \vdash B(p,q,r)^* \to I_{\Sigma_1}(\ulcorner \neg p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner). \tag{5}$$
$$\text{PA} \vdash \Sigma_1(\ulcorner q^* \urcorner) \leftrightarrow I_{\Sigma_1}(\ulcorner q^* \urcorner, \ulcorner q^* \urcorner). \tag{6}$$

Lemma For all φ_0, φ_1, and ψ,

$$\text{PA} \vdash I_{\Sigma_1}(\ulcorner \varphi_0 \urcorner, \ulcorner \psi \urcorner) \wedge I_{\Sigma_1}(\ulcorner \varphi_1 \urcorner, \ulcorner \psi \urcorner) \to I_{\Sigma_1}(\ulcorner \varphi_0 \vee \varphi_1 \urcorner, \ulcorner \psi \urcorner).$$

Proof Reason in PA. Assume that for some σ_0, $\Sigma_1(\sigma_0)$, PA $\vdash \varphi_0 \to \sigma_0$ and PA $\vdash \sigma_0 \to \psi$. And assume that for some σ_1, $\Sigma_1(\sigma_1)$, PA $\vdash \varphi_1 \to \sigma_1$ and PA $\vdash \sigma_1 \to \psi$. Then PA $\vdash \sigma_0 \vee \sigma_1 \to \psi$ and PA $\vdash \varphi_0 \vee \varphi_1 \to \sigma_0 \vee \sigma_1$. Since $\Sigma_1(\sigma_0 \vee \sigma_1)$, this concludes the proof. □

So, by the above lemma,

$$\text{PA} \vdash I_{\Sigma_1}(\ulcorner \neg p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner) \wedge I_{\Sigma_1}(\ulcorner p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner) \to I_{\Sigma_1}(\ulcorner q^* \urcorner, \ulcorner q^* \urcorner).$$

Now by (6) we have

$$\text{PA} \vdash \neg\Sigma_1(\ulcorner q^* \urcorner) \wedge I_{\Sigma_1}(\ulcorner \neg p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner) \to \neg I_{\Sigma_1}(\ulcorner p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner).$$

And thus by (5)

$$\text{PA} \vdash B(p, q, r)^* \to \neg I_{\Sigma_1}(\ulcorner p^* \wedge q^* \urcorner, \ulcorner q^* \urcorner). \tag{7}$$

So, combining (4) and (7), if we could express Σ_1-interpolability in ILM(S), by a formula $J(u, v)$ say, then the formula $J(p \wedge q, q)$ would be an interpolant for $A_s \to \neg B_r$.

Let us now return to the proof of Theorem 6.7. First let us see that ILM(S) $\vdash A_s \to \neg B_r$. On the one hand we have

$$\text{ILM(S)} \vdash \Sigma_1 s \wedge \Sigma_1 r \to \Sigma_1(s \vee r). \tag{8}$$

On the other hand we have

$$\text{ILM(S)} \vdash \Box(p \wedge q \to s) \wedge \Box(\neg p \wedge q \to r) \to \Box(q \to s \vee r)$$

and

$$\text{ILM(S)} \vdash \Box(s \to q) \wedge \Box(r \to q) \to \Box(s \vee r \to q).$$

Thus

$$\text{ILM(S)} \vdash A_s \wedge B_r \to \Box(q \leftrightarrow s \vee r). \tag{9}$$

Combining (8), (9), and S3 it follows that

$$\text{ILM(S)} \vdash A_s \wedge B_r \to \Sigma_1 q,$$

from which ILM(S) $\vdash A_s \to \neg B_r$ follows at once.

To show that no interpolant exists, the usual approach is to define two models M and M', based on frames for which ILM(S) is sound. Then one shows that for some $m \in M$ and $m' \in M'$ we have, on the one hand, $m \Vdash A_s$ and $m' \Vdash B_r$ and that, on the other hand, m and m' force exactly the same modal formulas that only use propositional variables among $\{p, q\}$. We will follow this idea. The necessary machinery is contained in the following definitions and lemmas.

Definition 6.8 (SL-bisimulation) Let M and M' be SL-models and P a set of propositional variables. A binary relation $Z \subseteq M \times M'$ is an SL-*bisimulation* with respect to P if the following conditions hold.

1. If wZw' then for each $p \in P$, $w \Vdash p \Leftrightarrow w' \Vdash' p$.
2. If wZw' then for all $\hat{w}, v, u \in M$, if $w(R \cup \bigcup_{x \in W} S_x)^* \hat{w}$ and $vS_{\hat{w}}u$, then there exist $\hat{w}', v', u' \in M'$ such that vZv', uZu', $w'(R \cup \bigcup_{x' \in W'} S_{x'})^* \hat{w}'$, and $v'S_{w'}u'$.
3. Same as 2 with M and M' interchanged.

Definition 6.9 (IL-bisimulation) Let M and M' be two models and P a set of propositional variables. A relation $Z \subseteq M \times M'$ is an IL-*bisimulation with respect to* P if the following conditions hold.

1. If wZw' then for each $p \in P$, $w \Vdash p \Leftrightarrow w' \Vdash' p$.
2. If wZw' and wRv then there exists v' such that vZv', $w'Rv'$ and for each u' with $v'(S_{w'})^*u'$ there exists some u such that uZu' and $v(S_w)^*u$.
3. Same as 2 with M and M' interchanged.

If $M = \langle W, R, S, \Vdash\rangle$ is an ILM(S)-model then with M^* we denote the model $\langle W', R', S^*, \Vdash'\rangle$, where S^* is the unique ternary relation on W such that for each $w \in W$ we have $(S^*)_w = (S_w)^*$. We clearly have $W = W'$ and $R = R'$. The next lemma shows that we also have $\Vdash = \Vdash'$.

Lemma 6.10 *Let $M = \langle W, R, S, \Vdash\rangle$ be a model and let $M^* = \langle W, R, S', \Vdash'\rangle$. Then for any* ILS*-formula A and any $w \in W$ we have $w \Vdash A$ if and only if $w \Vdash' A$.*

Proof Induction on A. The cases that A is a propositional variable and the cases for the Boolean connectives are trivial. So suppose that $A \equiv A_0 \rhd A_1$. Let $w \in W$ and suppose $w \Vdash' A$. We will show that $w \Vdash A$. Let $x \in W$ such that wRx and $x \Vdash A_0$. By (IH) we have $x \Vdash' A_0$ and thus for some y with $x(S'_w)^*y$ we have $y \Vdash' A_1$. Since $S'_w = (S_w)^*$ and $((S_w)^*)^* = (S_w)^*$ we also have $x(S'_w)^*y$. Also by (IH) we have $y \Vdash A_1$ and thus we are done. The case $w \Vdash A$ is even easier.

Now suppose $A \equiv \Sigma_1 B$. Let $w \in W$ and suppose $w \Vdash A$. We will show $w \Vdash' A$. Let $w' \in W$ such that $w(R \cup \bigcup_{u\in W} S'_u)^*w'$. Suppose $xS'_{w'}y$ and $x \Vdash' B$. Since for each $u \in W$ we have $S'_u = (S_u)^*$, we have $w(R \cup \bigcup_{u\in W} S_u)^*w'$. By (IH) we have $x \Vdash B$ and thus $y \Vdash B$. Again by (IH) we get $y \Vdash' B$ and we are done. The case $w \Vdash' A$ is even easier. □

Lemma 6.11 *Let P be a set of propositional variables. Let $M = \langle W, R, S, \Vdash\rangle$ and $M' = \langle W', R', S', \Vdash'\rangle$ be models and let Z be an* IL*-bisimulation (respectively,* SL*-bisimulation) between M and M' with respect to P. Then for any* IL*-formula (respectively,* SL*-formulas) I that only contains propositional variables among P and all $x \in M$ and $x' \in M'$ such that xZx', we have $x \Vdash I$ if and only if $x' \Vdash' I$.*

Proof Let Z be an IL-bisimulation as stated. We proceed with induction on I. The case that I is a propositional variable and the cases for the Boolean connectives are trivial. So suppose $I \equiv I_0 \rhd I_1$ and suppose that $x \Vdash I$. We will show that $x' \Vdash' I$. So suppose $x'R'y' \Vdash I_0$. Now there exists some $y \in M$ with xRy and yZy'. By (IH) we have $y \Vdash I_0$ and thus for some z with $y(S_x)^*z$ we have $z \Vdash I_1$. For some $z' \in M'$ we have $y'(S'_{x'})^*z'$ and zZz'. By (IH) we get $z' \Vdash' I_1$ and we are done. The case $x' \Vdash' I$ goes similarly.

Now suppose that Z is an SL-bisimulation. We proceed with induction on I. The case that I is a propositional variable and the cases for the Boolean connectives are trivial. So suppose $I \equiv \Sigma_1 J$ and suppose that $x \Vdash I$. We will show $x' \Vdash' I$. Suppose $x'(R \cup \bigcup_{t'\in W'} S_{t'})^*\hat{x}'$, $y'S'_{\hat{x}'}z'$ and $y' \Vdash' J$. There exist $\hat{x}$, y, and z with yZy', zZz' $x(R \cup \bigcup_{t\in W} S_t)^*\hat{x}$ and $yS_{\hat{x}}z$. By (IH) we get $y \Vdash J$ and thus $z \Vdash J$. Again by (IH) we have $z' \Vdash' J$ and we are done. □

Before we finish the proof of Theorem 6.7 we prove two more lemmas. In what follows we will write $D = D(p)$ to indicate all possible occurrences of a propositional variable p in the formula D. D does not necessarily contain p, and D might contain more variables different from p. The purpose of this notation is to indicate that with $D(A)$ we mean the formula that is the result of subsituting A for p in D.

Lemma 6.12 *For any formula $D = D(p)$ we have*

$$\mathsf{ILM(S)} \vdash \boxdot(A \leftrightarrow B) \to (D(A) \leftrightarrow D(B)).$$

Proof Induction on D. Suppose D is a proposition variable q. If $q \not\equiv p$ we have $D(A) \equiv D(B)$. If $q \equiv p$ then $D(A) \equiv A$ and $D(B) \equiv B$. Thus the claim is obvious in these cases. The cases for the propositional connectives are trivial.

Suppose $D(p) \equiv D_0(p) \rhd D_1(p)$. For $i \in \{0, 1\}$ we have by (IH), necessitation, and L1 that

$$\mathsf{ILM(S)} \vdash \boxdot(A \to B) \to \Box(D_i(A) \leftrightarrow D_i(B)).$$

Thus using J1 we get for $i \in \{0, 1\}$ that

$$\mathsf{ILM(S)} \vdash \boxdot(A \to B) \to (D_i(A) \rhd D_i(B)) \wedge (D_i(B) \rhd D_i(A)).$$

The claim now follows by J2.

Suppose $D(p) \equiv \Sigma_1 D'(p)$. By (IH), necessitation, and L1 we have

$$\mathsf{ILM(S)} \vdash \boxdot(A \to B) \to \Box(D'(A) \leftrightarrow D'(B)).$$

The claim now follows using S3. □

For a formula D and a propositional variable p, we say that p occurs modalized in D if any occurrences of p in D are under the scope of a $\rhd$ or Σ_1. (So in particular p might not occur in D at all.)

Lemma 6.13 *Let p be modalized in $D = D(p)$. Then for any formula A of the form $A_0 \rhd A_1$ or $\Sigma_1 A'$ we have*

$$\mathsf{ILM(S)} \vdash \Box\Box\bot \to (D(A) \leftrightarrow D(\top)).$$

Proof First notice that in any case we have $\mathsf{ILM(S)} \vdash \Box\bot \to A$ (if $A \equiv A_0 \rhd A_1$ then this follows by J1 and in case $A \equiv \Sigma_1 A'$ this follows using S3 and S5). And thus

$$\mathsf{ILM(S)} \vdash \Box\bot \to \boxdot(A \leftrightarrow \top). \tag{10}$$

We prove the lemma with induction on D. In case D is a propositional variable q then since p occurs modalized in D we have that $q \not\equiv p$ and thus the claim is clear. The cases for the propositional connectives are trivial.

Suppose $D(p) \equiv D_0(p) \rhd D_1(p)$. By (10) and Lemma 6.12 we have for $i \in \{0, 1\}$ that

$$\mathsf{ILM(S)} \vdash \Box\bot \to (D_i(A) \leftrightarrow D_i(\top)).$$

So for $i \in \{0, 1\}$ we have $\vdash \Box\Box\bot \to \Box(D_i(A) \leftrightarrow D_i(\top))$. Thus by J1 and J3 the claim follows.

Suppose $D(p) \equiv \Sigma_1 D'(p)$. By (10) and Lemma 6.12 we have

$$\mathsf{ILM(S)} \vdash \Box\bot \to (D'(A) \leftrightarrow D'(\top)).$$

Thus $\vdash \Box\Box\bot \to \Box(D'(A) \leftrightarrow D'(\top))$ and thus by S3 the claim follows. □

Proof of Theorem 6.7 Consider Figure 1. There are four models displayed there. In what follows we will use $\Vdash$ for each of the four forcing relations. Straight arrows indicate R relations, the wavy ones indicate S_w relations, and as we will argue below,

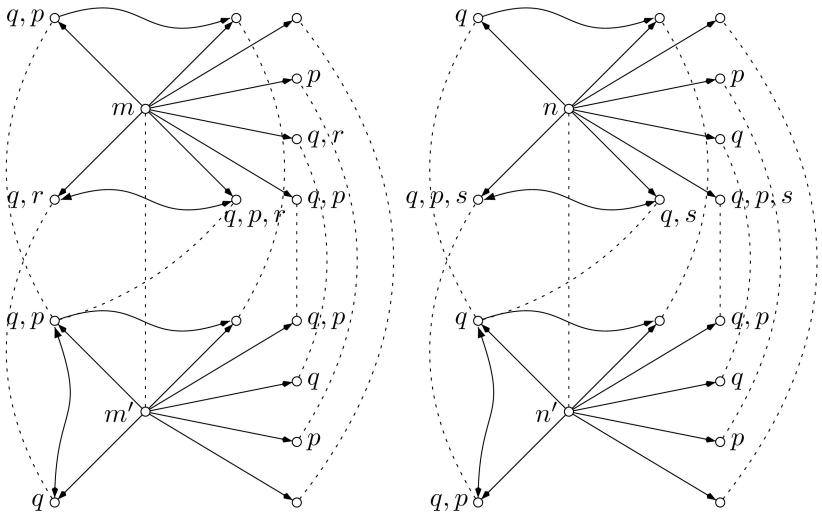

Figure 1 Two SL-bisimulations w.r.t. $\{p, q\}$.

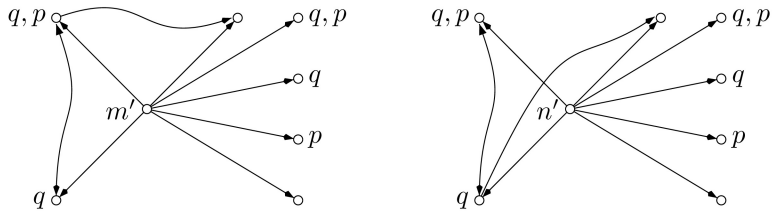

Figure 2 $(M')^* = (N')^*$.

the dotted lines are SL-bisimulations. One easily checks that all four models are ILM(S)-models (for clarity we have omitted the reflexive S relations) and that

$$m \Vdash B_r \text{ and } n \Vdash A_s.$$

Let $I = I(p, q)$ be a formula that at most contains the variables p and q. We will show the following.

$$m \Vdash I \Leftrightarrow n \Vdash I. \tag{11}$$

This would imply the theorem since if ILM(S) $\vdash A_r \to I \to \neg B_r$ then $n \Vdash I$ and therefore $m \Vdash I$. But then $m \Vdash \neg B_r$, a contradiction.

By Lemma 6.13 we see that on these models any ILS-formula is equivalent to a formula with no nested modalities, thus, in particular, equivalent to a Boolean combination of IL-formulas and SL-formulas. To show (11), it is therefore sufficient to show that m and n force the same SL-formulas as well as the same IL-formulas.

In Figure 2 we have depicted the models M' and N' again but have interchanged in N' the two leftmost nodes so that it is clear that we have

$$(M')^* = (N')^*.$$

So by Lemma 6.10 we have

$$m' \Vdash I \Leftrightarrow n' \Vdash I.$$

Thus by Lemma 6.11 it is sufficient to show that there exist IL-bisimulations with respect to $\{p, q\}$ between M and M' and between N and N', as well as SL-bisimulations with respect to $\{p, q\}$ between M and M' and between N and N', in both cases connecting m with m' and n with n'.

It is in order here to remark that M and N are not SL-bisimilar (Definition 6.8) and thus we really need Lemma 6.10 with these two models. After we have finished the proof of Theorem 6.7 we will argue that our little detour via Lemma 6.10 is in fact necessary.

Let us also briefly note that the four somewhat isolated nodes to the right of each of the models are there only to guarantee the existence of the IL-bisimulations. All other claims made about the models remain valid when these nodes are omitted.

The relation between M and M', call it Z, as indicated by the dotted lines, is an SL-bisimulation with respect to $\{p, q\}$. The definition of an SL-bisimulation in this case simplifies to the following two conditions.

$$x Z x' \text{ implies that } x \Vdash q \Leftrightarrow x' \Vdash q \text{ and } x \Vdash p \Leftrightarrow x' \Vdash p$$

and

$$\text{if } x' S_{m'} y' \text{ then for some } x, y \text{ we have } x S_m y,\ x Z x',\ \text{and } y Z y'.$$

And the other way around. The first claim is easily verified. For the second claim let us treat the case that x' is the left upper node in M' and y' is the middle upper node in M'. Then for x and y we can take the left upper node and the middle upper node in M, respectively. The other cases are verified similarly. The same holds for N and N'.

An IL-bisimulation between M and M', say Z, with respect to $\{p, q\}$, is constructed more easily. Between M and M' connect both center points, and for all other points $x \in M$ and $x' \in M'$, connect x with x' if and only if x and x' force the same propositional variables $\in \{p, q\}$. By construction $x Z x'$ implies that x and x' agree on $\{p, q\}$; thus, to show that Z is an IL-bisimulation it is sufficient to show that

$$\begin{aligned}&\text{if } m R x \text{ then for some } x' \text{ with } m' R x' \text{ we have that } x Z x' \text{ and}\\ &\qquad\text{for all } y' \text{ with } x'(S_{m'})^* y' \text{ there is some } y \text{ with } x(S_m)^* y \text{ and } y Z y'.\end{aligned}$$

And the other way around. For any choice of $x \in M$ we simply take x' to be one of the four rightmost nodes in M' for which $x Z x'$ (that is, the one that forces the same propositional variables among $\{p, q\}$ as x). Then a y' for which $x' S_{m'} y'$ can only be x' itself, and thus for y we can always take x. Similar for "the other way around." A completely similar argument shows that N and N' are IL-bisimilar. □

One might wonder whether the current proof can be simplified to the standard case: two models and a bisimulation between them. For starters, it is easy to see that there does not exist an SL-bisimulation that connects m with n (so incidentally this

shows that forcing the same SL-formulas does *not* imply the existence of an SL-bisimulation). But we can even make a stronger statement: for our particular counterexample any two models that do the trick are necessarily not bisimilar. For on the one hand one immediately sees that the above proof shows that $A_s \rightarrow \neg B_r$ is equally a counterexample for interpolation in SL as it is for ILM(S). On the other hand, in [11], the logic of Σ_1-interpolability is shown to satisfy the interpolation property. As this logic is an extension of SL, is also evaluated on Veltman models, is sound for ILM(S)-models, and the appropriate notion of bisimulation coincides with the notion of an SL-bisimulation, just bisimulations cannot do the job.

In particular this explains why we have chosen not to take the S_w relations transitive in the definition of an ILM model (Definition 2.5). The language of the logic of Σ_1-interpolability is not blind for this property in the sense that Lemma 6.10 does not hold there. And by the above we needed to exploit some difference between these logics.

7 Conclusion and Further Research

We have investigated the possibility of extending the language of interpretability logic with a modal operator which expresses Σ_1-ness. The primary reason for this investigation is the failure of interpolation in ILM, an important interpretability logic. We have formulated a modally complete and arithmetically complete logic ILM(S). Sadly this logic does not have interpolation either. Additionally, from the proofs it is immediate that the counterexample given in this paper is also a counterexample for interpolation in SL and, consequently, most likely also for HGL [8].

The reason for this seems to be a gap in expressive power which might be filled by the notion of Σ_1-interpolability. Ignatiev gave an arithmetically complete logic for Σ_1-interpolability and showed that his logic does have interpolation. In other words, extending SL with an operator for Σ_1-interpolability does give the interpolation property. However, recent investigations indicate that a combined logic of interpretability and Σ_1-interpolability is unlikely to have interpolation. More research is needed to say something more constructive in this direction.

Notes

1. This paper presents the main results from [9].

2. It is worthwhile to note that Beth definability is not an issue here; all interesting interpretability logics have the Beth definability property ([1] and de Jongh and Visser [7]).

3. In [9], ILM(S) was denoted by Σ_1ILM. See also the end of Section 3.

4. We contribute this theorem to Berarducci and Shavrukov, which is correct but somewhat incomplete. Indeed Berarducci and Shavrukov have independently shown ILM to be the interpretability logic of PA. By the Orey-Hájek characterization of interpretability this immediately gives that ILM is the logic of Π_1-conservativity for PA. In [16], Visser showed that ILM is not the interpretability logic of the theories $I\Sigma_n$, $n \geq 1$, but in Hájek and Montagna [10] it is shown that ILM is the logic of Π_1-conservativity for these theories as well.

5. In [17] it is shown that $\Box(p \leftrightarrow \Box q) \to (r \rhd s \to \Diamond r \wedge p \rhd s \wedge p)$ is a counterexample for interpolation. The proof works unmodified for $\Box(p \leftrightarrow \Box q) \to (r \rhd s \to r \wedge p \rhd s \wedge p)$, the original unpublished counterexample by Ignatiev, as well.

References

[1] Areces, C., E. Hoogland, and D. H. J. de Jongh,, "Interpolation, definability and fixed points in interpretability logics," pp. 35–58 in *Advances in Modal Logic. Vol. 2*, edited by M. Zakharyaschev, K. Segerberg, M. de Rijke, and H. Wansing, vol. 119 of *CSLI Lecture Notes*, CSLI Publications, Stanford, 2001. Selected papers from the 2nd International Workshop (AiML'98) Uppsala, 1998. Zbl 0979.00027. MR 1838243. 179, 193

[2] Beklemishev, L., "Notes on local reflection principles. The arithmetization of metamathematics," *Theoria*, vol. 63 (1997), pp. 139–46. MR 1730696. 185

[3] Berarducci, A., "The interpretability logic of Peano arithmetic," *The Journal of Symbolic Logic*, vol. 55 (1990), pp. 1059–89. Zbl 0725.03037. MR 1071315. 184, 187

[4] Boolos, G., *The Logic of Provability*, Cambridge University Press, Cambridge, 1993. Zbl 0891.03004. MR 1260008. 179, 183, 186

[5] de Jongh, D. H. J., and F. Veltman, "Modal completeness of ILW,", *Essays Dedicated to Johan van Benthem on the Occasion of His 50th Birthday*, edited by J. Gerbrandy, M. Marx, M. Rijke, and Y. Venema, Amsterdam University Press, Amsterdam, 1999. http://www.illc.uva.nl/j50/. 180

[6] de Jongh, D., and F. Veltman, "Provability logics for relative interpretability," pp. 31–42 in *Mathematical Logic (Proceedings of the 1988 Heyting Summer School)*, edited by P. Petkov, Plenum, New York, 1990. Zbl 0794.03026. MR 1083984. 182

[7] de Jongh, D., and A. Visser, "Explicit fixed points in interpretability logic," *Studia Logica*, vol. 50 (1991), pp. 39–49. Zbl 0744.03020. MR 1152779. 193

[8] Dzhaparidze, G., "The logic of arithmetical hierarchy," *Annals of Pure and Applied Logic*, vol. 66 (1994), pp. 89–112. Zbl 0804.03045. MR 1262432. 187, 193

[9] Goris, E., "Extending ILM with an operator for Σ_1-ness," Technical Report PP-2003-17, ILLC, Amsterdam, 2003. Preprint. 180, 182, 185, 186, 187, 193

[10] Hájek, P., and F. Montagna, "The logic of Π_1-conservativity," *Archive for Mathematical Logic*, vol. 30 (1990), pp. 113–23. Zbl 0713.03007. MR 1075648. 193

[11] Ignatiev, K. N., "The provability logic for Σ_1-interpolability," *Annals of Pure and Applied Logic*, vol. 64 (1993), pp. 1–25. Zbl 0802.03014. MR 1241249. 187, 193

[12] Japaridze, G., and D. de Jongh, "The logic of provability," pp. 475–546 in *Handbook of Proof Theory*, edited by S. R. Buss, vol. 137 of *Studies in Logic and the Foundations of Mathematics*, North-Holland, Amsterdam, 1998. Zbl 0915.03019. MR 1640331. 179, 180, 181, 182, 187

[13] Joosten, J. J., and E. Goris, "Modal Matters in Interpretability Logics," Technical Report LGPS-226, University of Utrecht, March 2004. Preprint. 180

[14] Shavrukov, V. Y., "The logic of relative interpretability over Peano arithmetic," Technical Report 5, Stekhlov Mathematical Institute, Moscow, 1988. (In Russian). 184, 187

[15] Solovay, R. M., "Provability interpretations of modal logic," *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287–304. Zbl 0352.02019. MR 0457153. 179

[16] Visser, A., "Interpretability logic," pp. 175–209 in *Mathematical Logic (Proceedings of the 1988 Summer School)*, edited by P. Petkov, Plenum, New York, 1990. Zbl 0793.03064. MR 1083994. 179, 193

[17] Visser, A., "An overview of interpretability logic," pp. 307–59 in *Advances in Modal Logic, Vol. 1 (Berlin, 1996)*, edited by M. Kracht, M. Rijke, and H. Wansing, vol. 87 of *CSLI Lecture Notes*, CSLI Publications, Stanford, 1998. Zbl 0915.03020. MR 1688529. 179, 180, 184, 194

Acknowledgments

I thank Dick de Jongh, Albert Visser, Joost J. Joosten, Maarten de Rijke, and the two anonymous referees for useful comments and suggestions.

City University of New York
365 5th Avenue
New York City NY 10016
evangoris@gmail.com