Interpretability in PRA

Marta Bílková, Dick de Jongh, Joost J. Joosten

Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodarenskou vezi 2, 182 07, Prague 8, Czech Republic

ILLC, Universiteit van Amsterdam, Plantage-Muidergracht 24, 1018 TV Amsterdam, The Netherlands

Departamento de Filosofía, Lógica y Filosofía de la Ciencia Universidad de Sevilla Camilo J. Cela, s/n. 41018 Sevilla, Spain

MSC: 03B45 03F30 03F45

Keywords: Interpretability Arithmetic Primitive recursive arithmetic Interpretability logic

Abstract

In this paper, we study \( \text{IL}(\text{PRA}) \), the interpretability logic of PRA. As PRA is neither an essentially reflexive theory nor finitely axiomatizable, the two known arithmetical completeness results do not apply to PRA: \( \text{IL}(\text{PRA}) \) is not \( \text{ILM} \) or \( \text{ILP} \). \( \text{IL}(\text{PRA}) \) does, of course, contain all the principles known to be part of \( \text{IL}(\text{All}) \), the interpretability logic of the principles common to all reasonable arithmetical theories. In this paper, we take two arithmetical properties of PRA and see what their consequences in the modal logic \( \text{IL}(\text{PRA}) \) are. These properties are reflected in the so-called Beklemeishev Principle \( B \) and Zambella's Principle \( Z \), neither of which is a part of \( \text{IL}(\text{All}) \). Both principles and their interrelation are submitted to a modal study. In particular, we prove a frame condition for \( B \). Moreover, we prove that \( Z \) follows from a restricted form of \( B \). Finally, we give an overview of the known relationships of \( \text{IL}(\text{PRA}) \) to important other interpretability principles.

1. Introduction

The notion of a relativized interpretation occurs in many places in mathematics and in mathematical logic. If a theory \( T \) interprets a theory \( S \), we shall write \( T \vdash S \), which then, roughly, means that there is a translation \( t \) from symbols in the language of \( S \) to formulas in the language of \( T \) such that any theorem of \( S \) becomes a theorem of \( T \) under the canonical extension of this translation to formulas. In the notion of interpretation that we are interested in, the logical structure of formulas has to be preserved under the translation. Thus, for example, \( (\varphi \lor \psi)^t = \varphi^t \lor \psi^t \) and in particular \( \bot^t = (\lor \varnothing)^t = \lor \varnothing = \bot \). We refer the reader to [17, 5, 15] for precise definitions and examples.

In this paper, we shall not go much into the technical details of interpretations. Rather, we are interested in the structural behavior of this notion of interpretability. In particular, we are interested in the structural behavior of interpretability on sentential extensions of a certain base theory \( T \). An easy example of such a structural property is the transitivity of interpretations:

\[
(T + \alpha \vdash T + \beta) \land (T + \beta \vdash T + \gamma) \rightarrow (T + \alpha \vdash T + \gamma).
\]

We can use so-called interpretability logics to capture, in a sense, the complete structural behavior of interpretability between sentential extensions of a certain base theory. We shall soon say a bit more on this. For now, it is important to note that, for a large collection of theories, the interpretability logic is known.

\( \star \) We thank Lev Beklemishev for his help and suggestions. Evan Goris did a thorough proofread of an early draft and suggested a simplification of the notion of B-simulation. We thank Albert Visser for fruitful discussions and challenges. We also thank Franco Montagna for his many contributions to the subject. Two unknown referees improved our paper considerably with their remarks and suggestions.

\( \star \) Corresponding author at: Institute of Computer Science, Academy of Sciences of the Czech Republic, Pod Vodarenskou vezi 2, 182 07, Prague 8, Czech Republic.

E-mail addresses: bilkova@cs.cas.cz, marta.bilkova@ff.cuni.cz (M. Bílková).

0168-0027/$ - see front matter © 2009 Elsevier B.V. All rights reserved.

doi:10.1016/j.apal.2009.05.012
We call a theory reflexive if it proves the consistency of any of its finite sub-theories (as sets of axioms). We call a theory essentially reflexive if any finite sentential extension of it is reflexive. It is easy to see that any theory with full induction, like Peano Arithmetic, is essentially reflexive. The interpretability logic of essentially reflexive theories was determined independently by Berarducci and Shavrukov [4, 13]. We shall encounter this logic below under the name of ILM. The principle \((A \supset B) \rightarrow (A \land C \supset B \lor C)\) is the particular feature of this system. It is called Montagna’s principle since it arose during the original discussions between Franco Montagna and Albert Visser about the modal principles underlying interpretability logic. It was known to Lindström and Švejdar in arithmetic disguise before.

It turns out that theories which are finitely axiomatizable and which contain a sufficient amount of arithmetic, have a different interpretability logic which is called ILP. In [17], the first proof was given.

For no theory that is neither finitely axiomatizable nor essentially reflexive, the interpretability logic is known. PRA is one such theory. In this paper, we shall make some first attempts to work out the interpretability logic of PRA.

As such, this paper also fits into a larger project. As pointed out above, different arithmetical theories have different interpretability logics. A question that is open since a long time concerns the logic of the core principles that pertain to all reasonable arithmetical theories — interpretability logics. In this paper, we shall not focus too much on the principles in the core logic. Rather shall we consider the interpretability behavior of PRA that is typical for this theory.

One such principle that is characteristic for PRA is Beklemishev’s principle that shall be studied closely in this paper. This principle exploits the fact that any theory which is an extension of PRA by \(\Sigma_2\) sentences is reflexive. We give a characterization of this principle in terms of the modal semantics for interpretability logics.

A topic that is closely related to interpretability logics, is that of \(\Pi_1\)-conservativity logics. A theory \(S\) is \(\Pi_1\) conservative over a theory \(T\) in the same language of arithmetic, we shall write \(S \vdash_{\Pi_1} T\) whenever \(S\) proves any \(\Pi_1\) theorem that is proven by \(T\). In symbols: \(T \vdash \pi \implies S \vdash \pi\) for any \(\pi \in \Pi_1\). It is easy to see that for any \(\Sigma_1\) sentence \(\sigma\), the following is a valid principle \(S \vdash_{\Pi_1} T \implies S + \sigma \vdash_{\Pi_1} T + \sigma\). This principle is the basis for Montagna’s principle for interpretability logic, and Beklemishev’s principle which is studied in this paper is a restriction of Montagna’s principle.

When \(T\) and \(S\) are both reflexive theories we have that \(S \vdash T \iff S \vdash_{\Pi_1} T\). This equivalence was exploited by Hájek and Montagna who were the first to show that the \(\Pi_1\)-conservativity logic of PA is ILM as well [9]. The observation about the equivalence is more generally important when looking at the repercussions of \(\Pi_1\)-conservativity principles on interpretability logics. In this paper, we shall consider Zambev’s principle for \(\Pi_1\)-conservativity logics and look at its repercussions for the interpretability logic of PRA. We shall show that Zambev does not add new information in the sense that its modal-logical consequences are already implied by Beklemishev’s principle.

It is remarkable that the notion of interpretability is, in a sense, less stable than that of \(\Pi_1\)-conservativity. Hájek and Montagna show that their results extend to all reasonable theories containing \(1 \Sigma_1\). This was strengthened by Beklemishev and Visser in [3]: all theories extending the parameter-free induction schema \(\Pi^1\) have the same \(\Pi_1\)-conservativity logic (ILM) whereas in this range the interpretability logics expose a diverse and wild behavior. Note though that PRA does not prove \(\Pi^1\), and, in fact, the \(\Pi_1\)-conservativity logic of PRA remains unknown.

A number of the results in this paper was first proved in [10].

2. Arithmetic

Let us first fix some arithmetical notation. We use modal symbols \(\Box, \Diamond, \supset\) both in modal and arithmetical statements, here we fix their arithmetical meaning. We write, for an arithmetical sentence \(\alpha\), for formal provability of \(\alpha\) in \(T\), \(\Box_{T, n} \alpha\) for formal provability of \(\alpha\) in \(T\) using only non-logical axioms with Gödel numbers \(\leq n\) and formulas of logical complexity \(\leq n\). Dually, \(\Diamond_{T, n} \alpha \equiv \neg \Box_{T, n} \neg \alpha\) means formalized consistency of \(\alpha\) over \(T\) (i.e. nonexistence of a proof of a contradiction from \(\alpha\)), while \(\Box_{T, n} \alpha\) means \(\neg \Diamond_{T, n} \neg \alpha\). For theories \(T\), \(S\) we use \(T \vdash S\) to denote formalized interpretability of \(S\) in \(T\). For arithmetical sentences \(\alpha, \beta, \beta \supset \beta\) means \(T + \alpha \supset T + \beta\). Similarly for theories \(T, S\), \(S \vdash_{\Pi_1} T\) denotes formalized \(\Pi_1\)-conservativity of \(T\) over \(S\) and for arithmetical sentences \(\alpha, \beta, \beta \supset_{\Pi_1} \beta\) means \(T + \alpha \vdash_{\Pi_1} T + \beta\).

2.1. What is PRA?

In the literature, there are many definitions of PRA. Probably the best known definition uses a language that contains a function symbol for every primitive recursive function. The axioms contain the defining equations of these functions. Moreover, there are induction axioms for each \(\Delta_0\)-formula in this enriched language.

Beklemishev has shown in [2] that PRA is in a strong sense equivalent (faithfully mutually interpretable) with \(\text{(EA)}_1^2\). Here, \(\text{(EA)}_1^2\) is the theory that is obtained by starting with \(\text{EA} (= 1 \Delta_0 + \exp)\) and iterating \(\omega\) many times’ \(\Pi_2\)-reflection. In symbols: \(\text{(EA)}_0^2 = \text{EA}\), and \(\text{(EA)}_{n+1}^2 = \text{RFN}_{\text{(EA)}_1}^1(\Pi_2)\).

\(^1\) Since PRA proves superexponentiation this is, in the case under study, equivalent to the restriction of axioms to those \(\leq n\).
In this paper, we shall use the definition:

\[ \text{PRA} := (\forall x \exists y)(\forall z)(\forall w)(\forall t)(\forall u)(\forall v)(\forall r)(\forall s)(\forall q)(\forall p)(\forall o)(\forall n)(\forall m)(\forall l)(\exists y)(\exists z)(\exists w)(\exists t)(\exists u)(\exists v)(\exists r)(\exists s)(\exists q)(\exists p)(\exists o)(\exists n)(\exists m)(\exists l) \]

Under this definition, the following lemma is immediate.

**Lemma 1.** Any r.e. extension of PRA by \( \Sigma^0_2 \) sentences is reflexive.

### 2.2. The Orey–Hájek characterizations

All theories that are mentioned here are supposed to be consistent and have a poly-time recognizable axiomatization. Orey and Hájek have given several equivalent conditions on theories which express that the one interprets the other. In this subsection, we shall briefly mention the one we shall need and refer to the literature for proofs.

**Lemma 2.** Whenever \( T \) is reflexive we have that

\[ T \vdash S \iff \forall x \; T \vdash \neg \Box_{S,x} \bot \]

Moreover in the presence of the totality of exponentiation this equivalence can be formalized.

\[ \vdash T \vdash S \iff \forall x \; \Box_{T,x} \neg \Box_{S,x} \bot \]

In [10] an overview is given of all the implications, corresponding requirements and necessary arguments regarding Orey–Hájek. In the above Lemma the \( \iff \) does not need the requirement of reflexivity and can actually be formalized in \( S^1_2 \).

For the other direction reflexivity is needed, and for its formalization, the totality of \( \exp \) as well.

Note that, using the above characterization, the prima facie \( S^1 \) notion of interpretability becomes \( \Pi^1_2 \).

### 3. Modal logics and semantics

Similarly as formalized provability can be captured by modal provability logic, we can use modal logic to reason about formalized interpretability. Modal logic proved to be an extremely useful tool to reason about such formalized phenomena since it can visualize their behavior using a simple language and an intuitive frame semantics. Perhaps the most significant point where modal logic shows its skills are completeness proofs — arithmetical completeness proofs are based on modal completeness proofs obtained by rather standard method of model theory of modal logics. For more on material contained in this section we refer to [17,10,8].

We will work with modal propositional language containing two modalities — a unary \( \Box \) modality for provability and a binary \( \Rightarrow \) modality for interpretability. Modal interpretability formulas are defined as follows:

\[ \mathcal{A} ::= p | \bot | (\mathcal{A} \land \mathcal{B}) | (\mathcal{A} \Rightarrow \mathcal{B}) | (\Box \mathcal{A}) | (\Diamond \mathcal{A}) \]

We will use standard abbreviations \( \Diamond, \lor, \neg, \top, \Leftarrow, \Rightarrow \), and we write \( \mathcal{A} \equiv \mathcal{B} \) instead of \( (\mathcal{A} \Rightarrow \mathcal{B}) \land (\mathcal{B} \Rightarrow \mathcal{A}) \). We shall often omit brackets writing formulas. We say that \( \neg, \Box, \land \) bind equally strongly, they bind stronger then equally strong binding \( \lor \) and \( \top \) which in turn bind stronger then \( \Rightarrow, \Leftarrow \). The weakest binding connectives are \( \rightarrow \) and \( \Leftarrow \).

An arithmetical interpretation of modal formulas is given by arithmetical realizations: for an arithmetical theory \( T \), an arithmetical \( T \)-realization is a map \( * \) sending propositional variables \( p \) to arithmetical sentences \( p^* \). It is extended to interpretability modal formulas as follows: first \( \ast \) commutes with all boolean connectives. Moreover \( (\Box^* \mathcal{A})^* = \Box^* \mathcal{A}^* \) and \( (\mathcal{A} \Rightarrow^* \mathcal{B})^* = \mathcal{A}^* \Rightarrow^* \mathcal{B}^* \), i.e. \( * \) translates modal operators to formalized provability and interpretability over \( T \) respectively.

An interpretability principle of an arithmetical theory \( T \) is a modal formula \( \mathcal{A} \) such that \( \forall \mathcal{A} \vdash \mathcal{A}^* \). The interpretability logic of a theory \( T \), denoted \( \mathsf{IL}(T) \), is then the set of all the interpretability principles of \( T \).

#### 3.1. The logic \( \mathsf{IL} \)

The logic \( \mathsf{IL} \) is in a sense the core interpretability logic — it is a (proper) part of the interpretability logic of any reasonable arithmetical theory: \( \mathsf{IL} \subseteq \mathsf{IL}(T) \). It captures the basic structural behavior of interpretability.

\( \mathsf{IL} \) is defined as the smallest set of formulas containing all propositional tautologies, all instantiations of the following schemata, and is closed under the Necessitation and Modus Ponens rules:

\[
\begin{align*}
\mathsf{L1} & \quad \Box (A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B) \\
\mathsf{L2} & \quad \Box A \rightarrow \Box \Box A \\
\mathsf{L3} & \quad \Box (\Box A \rightarrow A) \rightarrow \Box A \\
\mathsf{J1} & \quad \Box (A \rightarrow B) \rightarrow A \Rightarrow B \\
\mathsf{J2} & \quad (A \Rightarrow B) \land (B \Rightarrow C) \rightarrow A \Rightarrow C \\
\mathsf{J3} & \quad (A \Rightarrow C) \land (B \Rightarrow C) \rightarrow A \lor B \Rightarrow C \\
\mathsf{J4} & \quad A \Rightarrow B \rightarrow (\Box A \Rightarrow \Box B) \\
\mathsf{J5} & \quad \Diamond A \Rightarrow A \\
\end{align*}
\]
Note that the part of \( \text{IL} \) not containing the \( \triangleright \) modality is the well-known Gödel–Löb provability logic \( \text{GL} \), axiomatized by the first three schemata. It is easy to show that \( \square \) can be defined in terms of \( \triangleright \) modality: \( \vdash_{\text{IL}} \square A \iff \neg \triangleright \neg A \).

More interpretability logics are obtained extending \( \text{IL} \) by new interpretability principles. Some of such principles are listed below:

\[
\begin{align*}
W & A \triangleright B \rightarrow A \triangleright B \land \neg \neg A \\
W^* & A \triangleright B \rightarrow B \land \square C \triangleright B \land \square C \land \neg \neg A \\
M_0 & A \triangleright \triangleright \triangleright \triangleright \triangleleft A \\
M & A \triangleright B \rightarrow A \land \square C \triangleright B \land \square C \\
P & A \triangleright B \rightarrow \square (A \triangleright B) \\
R & A \triangleright B \rightarrow \neg (A \triangleright \neg C) \triangleright B \land \square C \\
R^* & A \triangleright B \rightarrow \neg (A \triangleright \neg C) \triangleright B \land \square C \land \square \neg \neg A
\end{align*}
\]

All of these principles are in \( \text{IL}(\forall) \) except the principles \( M \) and \( P \) which were mentioned above already. For an overview, see [17,8]. For the last word on \( \text{IL}(\forall) \) see [11].

For a set of principles we denote \( \text{IL}(X) \) the logic extending \( \text{IL} \) with schemata from \( X \).

There are some results considering arithmetical completeness of interpretability logics: it was shown in [4,13] that the interpretability logic of an essentially reflexive theory (as e.g. PA) is \( \text{IL}(\forall) \).

For finitely axiomatizable theories containing \( \Pi^1_1 \)-conservativity of interpretability more directly is \( (A \triangleright \bigotimes B) \rightarrow \square (A \rightarrow \bigotimes B) \).

3.2. Modal semantics

Modal frame semantics of interpretability logics is based on \( \text{GL} \)-frames extended with a ternary accessibility relation interpreting the binary \( \triangleright \) modality. The ternary relation is however given by a set of binary relations indexed by the nodes:

**Definition 1.** An \( \text{IL} \)-frame (a Veltman frame) is a triple \( \langle W, R, S \rangle \) where \( W \) is a nonempty universe, \( R \) is a binary relation on \( W \), and \( S \) is a set of binary relations on \( W \) indexed by elements of \( W \) such that

1. \( R \) is transitive and conversely well-founded
2. \( yS_\langle z \Rightarrow xRy \land xRz \)
3. \( xRy \Rightarrow yS_\langle y \)
4. \( xRy \Rightarrow yS_\langle z \)
5. \( uS_\langle vS_\langle u \Rightarrow uS_\langle v \)

An \( \text{IL} \)-model is a quadruple \( \langle W, R, S, \models \rangle \) where \( \langle W, R, S \rangle \) is a \( \text{IL} \)-frame and \( \models \) is a subset of \( W \times \text{Prop} \), extending to boolean formulas as usually and to modal formulas as follows:

\[
\begin{align*}
w \models \square A \iff & \exists u (u \models Rv \land v \models A) \\
w \models A \models B & \iff \forall u (u \models Ru \land u \models A \Rightarrow \exists v (u \models Ru \land v \models B))
\end{align*}
\]

We adopt standard definitions of validity of a modal formula in a model and in a frame. Moreover, let \( X \) be a scheme of interpretability logic. We say that a formula \( C \) in first or higher order logic is a frame condition for \( X \) if, for each frame \( F \),

\( F \models C \iff F \models X \).

Let us list some known frame conditions (to be read universally quantified):

\[
\begin{align*}
M & xRyS_\langle zRu \Rightarrow yRu \\
M_0 & xRyRzS_\langle uRv \Rightarrow yRv \\
P & xRyRzS_\langle u \Rightarrow yRu \land zS_\langle u \\
W & (S_\langle u ; R) \text{ is conversely well-founded} \\
R & xRyRzS_\langle uRv \Rightarrow zS_\langle v
\end{align*}
\]

We have the following completeness results: \( \text{IL} \) is sound and complete w.r.t. (finite) \( \text{IL} \) frames, \( \text{ILP} \) is complete w.r.t. (finite) \( \text{ILP} \) frames (all in [6]), \( \text{ILW} \) is complete w.r.t. (finite) \( \text{ILW} \) frames ([7], see also [8]), \( \text{ILM} \) is complete w.r.t. (finite) \( \text{ILM} \) frames (in [6], also in [4]).

4. Beklemishev’s principle

It is possible to write down a valid principle specific for the interpretability logic of PRA. This was first done by Beklemishev (see [17]). Beklemishev’s principle \( B \) exploits the fact that any finite \( \Sigma^1_2 \)-extension of PRA is reflexive, together with the fact that we have a good Orey–Hájek characterization for reflexive theories.
It turns out to be possible to define a class of modal formulae which are under any arithmetical realization provably $\Sigma_2$ in PRA. These are called essentially $\Sigma_2$-formulas, we write ES$_2$. Let us start by defining this class and some related classes.

The idea behind this definition is as follows. It is clear that each modal formula that starts with a $\square$ will become under any arithmetical realization a $\Sigma_1$ formula. Likewise, taking Lemma 2 into account, we see that any formula of the form $A \rightarrow B$ where $A$ is $\Sigma_2$, will be under any arithmetical realization of complexity $\Pi_2$ and hence, $\neg(A \rightarrow B)$ will again be $\Sigma_2$. Note that we are here only formulating sufficient conditions. It turns out to be rather tough to show these classes actually cover, up to provably equivalent, all formulae in the intended complexity class.

The class BS$_1$ denotes the formulae that are boolean combinations of $\Sigma_1$ formulae ad thus certainly $\Delta_2$. Likewise, ES$_3$ and ES$_4$, stands for those modal formulae that are under any arithmetical realization always $\Sigma_3$ or $\Sigma_4$ respectively.

In our definition, $A$ will stand for the set of all modal interpretability formulae.

$$\begin{align*}
BS_1 & ::= \square A | \neg BS_1 | BS_1 \land BS_1 | BS_1 \lor BS_1 \\
BS_2 & ::= \square A | \neg BS_2 | BS_2 \land BS_2 | BS_2 \lor BS_2 | \neg(ES_2 \rightarrow A) \\
BS_3 & ::= \square A | \neg BS_3 | BS_3 \land BS_3 | BS_3 \lor BS_3 | A \rightarrow A \\
BS_4 & ::= \square A | \neg BS_4 | BS_4 \land BS_4 | BS_4 \lor BS_4 | A \rightarrow A
\end{align*}$$

For $n \geq 4$ we set $ES_n := ES_4$. We can now formulate Beklemishev’s principle $B$.

$$B ::= A \rightarrow B \rightarrow A \land \square C \rightarrow B \land \square C \quad \text{for } A \in ES_2$$

Note that $B$ is just Montagna’s principle $M$ restricted to $ES_2$-formulae.

**Lemma 3.** ILB $\vdash B'$, where $B' : A \rightarrow B \rightarrow A \land \square C \rightarrow B \land \square C$ with $A \in ES_2$ and $C$ a CNF (a conjunction of disjunctions) of boxed formulas.

**Proof.** Easy. $\square$

5. Arithmetical soundness of $B$

By **Lemma 1** we know that PRA + $\sigma$ is reflexive for any $\Sigma_2$(PRA)-sentence $\sigma$. Thus, we get by Orey–Hájek that

$$\text{PRA} \vdash \sigma \rightarrow (\forall x \square \text{PRA} \sigma \rightarrow \diamond \text{PRA}_x \psi). \quad (1)$$

Consequently, for $\sigma \in \Sigma_2$(PRA), $\neg(\sigma \rightarrow (\forall x \square \text{PRA} \sigma \rightarrow \diamond \text{PRA}_x \psi) \in \Sigma_2$(PRA) and we see that, indeed, $\forall A \in ES_2 \forall \ast A^\ast \in \Sigma_2$(PRA). This enables us to prove the arithmetical soundness of $B$.

**Theorem 1.** For any formulas $B$ and $C$ we have that $\forall A \in ES_2 \forall \ast A \in PRA \vdash (A \rightarrow B \rightarrow A \land \square C \rightarrow B \land \square C)^\ast$.

**Proof.** For some $A \in ES_2$ and arbitrary $B$ and $C$, we consider some realization $\ast$ and let $\alpha := A^\ast$, $\beta := B^\ast$ and $\gamma := C^\ast$. We reason in PRA and assume $\alpha \rightarrow PRA \beta$. As $\alpha$ is $\Sigma_2(PRA)$, we get by (1) that

$$\forall x \square \text{PRA}(\alpha \rightarrow \diamond \text{PRA}_x \beta). \quad (2)$$

We now consider $n$ large enough (dependent on $\gamma$) such that

$$\square \text{PRA}(\diamond \text{PRA}_x \gamma) \rightarrow \square \text{PRA}_x \text{PRA}_n(\gamma) . \quad (3)$$

From general observations we have that, for large enough $n$,

$$\square \text{PRA}_n(\delta \rightarrow \neg \epsilon) \land \square \text{PRA}_n \delta \rightarrow \square \text{PRA}_n \neg \epsilon,$$

whence

$$\diamond \text{PRA}_n \epsilon \land \square \text{PRA}_n \delta \rightarrow \diamond \text{PRA}_n (\delta \land \epsilon) \quad (4)$$

Combining (2), (3), and using (4), we see that for any $n$, $\square(\alpha \land \square \gamma \rightarrow \diamond \text{PRA}_n (\beta \land \square \gamma))$. Clearly, $\alpha \land \square \gamma$ is still a $\Sigma_2$(PRA)-sentence. Again by (1) we get $\alpha \land \square \gamma \rightarrow \beta \land \square \gamma$. $\square$

Let $M^{ES_n}$ be the schema $A \rightarrow B \rightarrow A \land \square C \rightarrow B \land \square C$ with $A \in ES_n$. **Theorem 1** can be generalized using results of [1] to the theory $I \Sigma_n^R$, which is Robinson’s arithmetic Q plus the $I \Sigma_n$ induction rule, for $n = 1, 2, 3$ as follows:

**Theorem 2.** IL($I \Sigma_n^R$) $\vdash M^{ES_{n+1}}$ for $n = 1, 2, 3$.

---

2 Actually, this observation is not necessary as we use the direction in the Orey–Hájek Characterization that does not rely on the reflexivity.
6. A frame condition for B

Let us first fix some notation. If \( C \) is a finite set, we write \( xRc \) as short for \( \bigwedge_{c \in C} xRc \). Similar conventions hold for the other relations. The \( A \)-critical cone of \( x \), \( C^A_x \) is in this section defined as \( C^A_x := \{ y \mid xRy \land \forall z (yzz \rightarrow z \neq A) \} \).

By \( y^\uparrow \) we denote the set of worlds that lie above \( y \) w.r.t. the \( R \) relation. That is, \( x^\uparrow := \{ y \mid xRy \} \). With \( yS^\uparrow \) we denote the set of those \( z \) for which \( ySz \).

We will consider frames both as modal models without a valuation and as structures for first- (or sometimes second) order logic. We say that a model \( M \) is based on a frame \( F \) if \( F \) is precisely \( M \) with the \( \models \) relation left out.

In this subsection we give the frame condition of Beklemishev’s principle. Our frame condition holds on the class of finite frames. At first sight, the condition might seem a bit awkward. On second sight it is just the frame condition of \( M \) with some simulation built in. First we approximate the class \( ES_2 \) by stages.

**Definition 2.**

\[
\begin{align*}
ES_2^0 & := BS_1 \\
ES_2^{n+1} & := ES_2^n \land ES_2^{n+1} \land ES_2^{n+1} \lor ES_2^{n+1} \land \neg(ES_2^n \triangleright A)
\end{align*}
\]

It is clear that \( ES_2 = \bigcup_i ES_2^i \). We now define some first order formulas \( \delta_i(b, u) \) that say that two nodes \( b \) and \( u \) in a frame look alike. The larger \( i \) is, the more the two points look alike. We use the letter \( \delta \) as to hint at a simulation.

**Definition 3.**

\[
\begin{align*}
\delta_0(b, u) & := b \uparrow = u \uparrow \\
\delta_{n+1}(b, u) & := \delta_{n}(b, u) \land \\
& \forall c (bRc \rightarrow \exists c’ (uRc’ \land \delta_{n}(c, c’) \land c’S_{c’} \uparrow \subseteq cS_{b} \uparrow))
\end{align*}
\]

By induction on \( n \) we easily see that \( \forall n F \models \delta_n(b, u) \) for all frames \( F \) and all \( b \in F \). For \( i \geq 1 \) the relation \( \delta_i(b, u) \) is in general not symmetric. However it is not hard to see that the \( \delta_i \) are transitive and reflexive.

**Lemma 4.** Let \( F \) be a model. For all \( n \) we have the following. If \( \models \delta_n(b, u) \), then \( b \models A \Rightarrow u \models A \) for all \( A \in ES_2^n \).

**Proof.** We proceed by induction on \( n \). If \( n = 0 \), \( A \in ES_2^0 \) can be written as \( \bigwedge_i (\Box A_i \land \bigwedge_j \diamond A_j) \). Clearly, if \( b \uparrow = u \uparrow \) then \( b \models A \Rightarrow u \models A \).

Now consider \( A \in ES_2^{n+1} \) and \( b \) and \( u \) such that \( F \models \delta_{n+1}(b, u) \). We can write

\[
A = \bigwedge_i \left( A_{0i} \land \bigwedge_{j \neq 0} \neg (A_{ij} \triangleright B_{ij}) \right)
\]

with \( A_{ij} \) in \( ES_2^0 \). If \( b \models A \), then for some \( i, b \models A_{0i} \land \bigwedge_{j \neq 0} \neg (A_{ij} \triangleright B_{ij}) \). As \( \delta_{n+1}(b, u) \rightarrow \delta_{n}(b, u) \), and by the induction hypothesis we see that \( b \models A_{0i} \). So, we only need to see that \( u \models \neg (A_{ij} \triangleright B_{ij}) \) for \( j \neq 0 \). As \( b \models \neg (A_{ij} \triangleright B_{ij}) \), for some \( c \in C^A_{b} \) we have \( c \models A_{ij} \). By \( \delta_{n+1}(b, u) \) we find a \( c’ \) such that \( uRc’ \) and \( c’S_{c’} \uparrow \subseteq cS_{b} \uparrow \) (thus \( cS_{b}c’ \)). This guarantees that \( c’ \in C^A_{u} \). Moreover we know that \( \delta_{n}(c, c’) \), thus by the induction hypothesis, as \( c \models A_{ij} \), we get that \( c’ \models A_{ij} \). Consequently \( u \models \neg (A_{ij} \triangleright B_{ij}) \). \( \square \)

**Lemma 5.** Let \( F \) be a finite frame. For all \( i \), and any \( b \in F \), there is a valuation \( V^b_i \) on \( F \) and a formula \( A^b_i \in ES_2^i \) such that \( F \models \delta_i(b, u) \Leftrightarrow u \models A^b_i \).

**Proof.** The proof proceeds by induction on \( i \). First consider the basis case, that is, \( i = 0 \). Let \( b \uparrow \) be given by the finite set \( \{ x_j \}_{j \in \mathcal{E}} \). We define

\[
\begin{align*}
y \models p_j & \Leftrightarrow y = x_j \\
y \models r & \Leftrightarrow bRy.
\end{align*}
\]

Let \( A^0_i \) be \( \Box r \land \bigwedge_j \diamond p_j \). It is now obvious that \( u \models A_0 \Leftrightarrow u \uparrow = b \uparrow \).

For the inductive step, we fix some \( b \) and reason as follows. First, let \( V^b_i \) and \( A^b_i \) be given by the induction hypothesis such that \( u \models A^b_i \Leftrightarrow F \models \delta_i(b, u) \). We do not specify the variables in \( A_i \), but we suppose they do not coincide with any of the ones mentioned below. Let \( b \uparrow = \{ x_j \}_{j \in \mathcal{E}} \). The induction hypothesis gives us sentences \( A^i_j \) (no sharing of variables) and valuations \( V^b_j \) such that \( F, u \models A^i_j \Leftrightarrow F \models \delta_i(x_j, u) \).

Let \( \{ q_j \}_{j \in \mathcal{E}} \) be a set of fresh variables. \( V^b_{i+1} \) will be \( V^b_i \) and \( V^i_j \) on the old variables. For the \( \{ q_j \}_{j \in \mathcal{E}} \) we define \( V^b_{i+1} \) to act as follows:

\[
y \models q_j \Leftrightarrow y \not\in xS^\uparrow_u.
\]

Moreover we define

\[
A^{b+1} := A^b_i \land \bigwedge_j \neg (A^i_j \triangleright q_j).
\]

Now we will see that under the new valuation \( V^b_{i+1} \).
(i) \( u \models A_{i+1}^b \Rightarrow F \models \delta_{i+1}(b, u) \),
(ii) \( F \models \delta_{i+1}(b, u) \Rightarrow u \models A_{i+1}^b \).

For (i) we reason as follows. Suppose \( u \models A_{i+1}^b \), then also \( u \models A_i^b \) and thus \( F \models \delta_i(b, u) \). It remains to show that

\[ F \models \forall c (b Rc \rightarrow \exists c' (u Rc' \land \delta_i(c, c') \land cSb_c' \land c'S_c b' \subseteq cS_{b'})) \]

To this purpose we consider and fix some \( x_j \) in \( b \). As \( u \models A_{i+1}^b \), we get that \( u \models \neg (A_j^b \models q) \). Thus, for some \( c' \in C_i^b \), \( c' \models A_j^i \). Clearly \( c' \models \neg q \), whence \( x_j A c' \). Also \( Vt (c'S_{b_c'} \Rightarrow y \models \neg q) \) which, by the definition of \( Vt_{b_j} \), translates to \( cS_{b'} \subseteq x_j A c' \). Clearly also \( u Rc' \). By \( c' \models A_j^i \) and the induction hypothesis we get that \( \delta_i(x_j, c') \). Indeed we see that \( F \models \delta_{i+1}(b, u) \).

For (ii) we reason as follows. As \( F \models \delta_{i+1}(b, u) \), also \( F \models \delta_i(b, u) \) and by the induction hypothesis, \( u \models A_i^b \). It remains to show that \( u \models \neg (A_j^b \models q) \) for any \( j \). So, let us fix some \( j \). Then, by the second part of the \( \delta_{i+1} \) requirement we find a \( c' \) such that

\[ u Rc' \land \delta_i(x_j, c') \land x_j A c' \land c'S_{b'} \subseteq x_j A c' \].

Now, \( u Rc' \land \delta_i(x_j, c') \land x_j A c' \subseteq x_j A c' \) gives us that \( c' \in C_i^b \). By \( \delta_i(x_j, c') \) and the induction hypothesis we get that \( c' \models A_j^i \). Thus indeed \( u \models \neg (A_j^b \models q) \).

Note that, in the proof of this lemma, we have only used conjunctions to construct the formulas \( A_i^b \).

**Definition 4.** For every \( i \) we define the frame condition \( C_i \) to be

\[ \forall a, b \ (aRb \rightarrow \exists u ((bS_u \land \delta_i(b, u) \land \forall d, e \ (uS_dRe \rightarrow bRe))). \]

**Lemma 6.** Let \( F \) be a finite frame. For all \( i \), we have that

\[ F \models \delta_i \land \forall A \models B \rightarrow A \land \square C \lor B \land \square C, \]

if and only if \( F \models C_i \).

**Proof.** First suppose that \( F \models C_i \) and that \( a \models A \models B \) for some \( A \in C_i^b \) and some valuation on \( F \). We will show that \( a \models A \land \square C \models B \land \square C \) for any \( C \). Consider therefore some \( b \) with \( aRb \) and \( b \models A \land \square C \). The \( C_i \) condition provides us with \( u \) such that

\[ bS_u \land \delta_i(b, u) \land \forall d, e \ (uS_dRe \rightarrow bRe) \]

As \( F \models \delta_i(b, u) \), we get by Lemma 4 that \( u \models A \). Thus, as \( aRu \) and \( a \models A \models B \), we know that there is some \( d \) with \( uS_d \) and \( d \models B \). If now \( dRe \), by \( (*) \), also \( bRe \) and hence \( e \models C \). Thus, \( d \models B \land \square C \). Clearly \( bS_u \) and thus \( a \models A \land \square C \models B \land \square C \).

For the opposite direction we reason as follows. Suppose that \( F \models C_i \). Thus, we can find \( a, b \) with

\[ aRb \land \forall u ((bS_u \land \delta_i(b, u) \land \exists e \ (uS_dRe \land \neg bRe))). \]

By Lemma 5 we can find a valuation \( V_t^b \) and a sentence \( A_i^b \in C_i^b \) such that \( u \models A_i^b \) \( \rightharpoonup F \models \delta_i(b, u) \). Let \( q \) and \( s \) be fresh variables. Moreover, let \( D \) be the following set.

\[ D := \{d \in F \mid bS_u \land \neg bRe \text{ for some } e\}. \]

We define a valuation \( V \) that is an extension of \( V_t^b \) by stipulating that

\[ y \models q \iff (y \in D) \lor \neg (bS_u y), \]

\[ y \models s \iff bRy. \]

We now see that

(i) \( a \models A_i^b \models q \),
(ii) \( a \models \neg (A_i^b \land \square D \models q \land \square \ell) \).

For (i) we reason as follows. Suppose that \( aRb' \) and \( b' \models A_i^b \). If \( \neg (bS_u b') \), then \( b' \models q \) and we are done. So, we consider the case in which \( bS_b \). As \( \delta_i(b, b') \), \( (** \) ) yields now us a \( d \in D \) such that \( bS_u d \). Clearly \( bS_u d \) and thus, by definition, \( d \models q \).

To see (ii) we notice that \( b \models A_i^b \land \square D \). But if \( bS_u y \) and \( y \models q \), by definition \( y \in D \) and thus \( y \models \neg \ell \). Thus \( b \in C_i^b \) and \( a \models \neg (A_i^b \land \square D \models q \land \ell) \).

The following theorem is now an immediate corollary of the above reasoning.

**Theorem 3.** A finite frame \( F \) validates all instances of Beklemishev’s principle if and only if \( \forall i \ F \models C_i \).

**Definition 5.** Let \( B_i \) be the principle \( A \models B \rightarrow A \land \square C \rightarrow B \land \square C \) for \( A \in C_i^b \).

**Corollary.** For a finite frame we have \( F \models B_i \leftrightarrow F \models C_i \).
For the class of finite frames, we can get rid of the universal quantification in the frame condition of Beklemishev’s principle. Remember that depth(x), the depth of a point x, is the length of the longest chain of R-successors starting in x.

**Lemma 7.** If \( \delta_n(x, x') \), then depth(x) = depth(x').

**Proof.** \( \delta_n(x, x') \Rightarrow \delta_0(x, x') \Rightarrow x^+ = x'^+ \). □

**Lemma 8.** If \( \delta_n(x, x') \) & depth(x) ≤ n, then \( \delta_m(x, x') \) for all m.

**Proof.** The proof goes by induction on n. For n = 0, the result is clear. So, we consider some x, x' with \( \delta_{n+1}(x, x') \) & depth(x) ≤ n + 1. We are done if we can show \( \delta_{m+1}(x, x') \) for m ≥ n + 1.

This, we prove by a subsidiary induction on m. The basis is trivial. For the inductive step, we assume \( \delta_m(x, x') \) for some m ≥ n + 1 and set out to prove \( \delta_{m+1}(x, x') \), that is

\[
\delta_m(x, x') \land \forall y (xRy \rightarrow \exists y' (yS\alpha y' \land \delta_m(y, y') \land y'S\beta y' \subseteq yS\beta y' ))
\]

The first conjunct is precisely the induction hypothesis. For the second conjunct we reason as follows. As m ≥ n + 1, certainly \( \delta_{n+1}(x, x') \). We consider y with xRy. By \( \delta_{n+1}(x, x') \), we find a y' with

\[
yS\alpha y' \land \delta_m(y, y') \land y'S\beta y' \subseteq yS\beta y' .
\]

As xRy and depth(x) ≤ n + 1, we see depth(y) ≤ n. Hence by the main induction, we get that \( \delta_m(y, y') \) and we are done. □

**Definition 6.** A B-simulation on a frame is a binary relation \( \delta \) for which the following holds.

1. \( \delta(x, x') \rightarrow x^+ = x'^+ \)
2. \( \delta(x, x') \land xRy \rightarrow \exists y' (yS\alpha y' \land \delta(y, y') \land y'S\beta y' \subseteq yS\beta y' ) \)

If F is a finite frame that satisfies \( C_i \) for all i, we can consider \( \bigcap_{i<\omega} C_i \). This will certainly be a B-simulation.

**Definition 7.** The frame condition \( C_0 \) is defined as follows. F \( \models C_0 \) if and only if there is a B-simulation \( \delta \) on F such that for all x and y,

\[
xRy \rightarrow \exists y' (yS\alpha y' \land \delta(y, y') \land \forall d, e (y'S\alpha dRe \rightarrow yS\beta e))
\]

An immediate consequence of **Lemma 8** is the following theorem.

**Theorem 4.** For F a finite frame, we have

\[
F \models B \iff F \models C_0 .
\]

Note that the M-frame condition can be seen as a special case of the frame condition of B: we demand that \( \delta \) be the identity relation.

It is not hard to see that the frame condition of \( M_0 \) follows from \( C_0 \). And indeed, \( \mathbf{IL}_0 \vdash M_0 \) as \( \Diamond A \in \mathbf{ES}_2 \) and \( A \vdash B \rightarrow \Diamond A \vdash B \). Actually, we have that \( \mathbf{IL}_0 \vdash M_0 \).

7. Beklemishev and Zambella

Zambella proved in [18] a fact concerning \( \Pi_1 \)-consequences of theories with a \( \Pi_2 \) axiomatization. As we shall see, his result has some repercussions on the study of the interpretability logic of PRA.

**Lemma 9** (Zambella). Let T and S be two theories axiomatized by \( \Pi_2 \)-axioms. If T and S have the same \( \Pi_1 \)-consequences then \( T + S \) has no more \( \Pi_1 \)-consequences than T or S.

In [18], Zambella gave a model-theoretic proof of this lemma. As was sketched by Mints (see [3]), also a finitary proof based on Herbrand’s theorem can be given. This proof can certainly be formalized in the presence of the superexponentiation function, thus it yields a principle for the \( \Pi_1 \)-conservativity logic of \( \Pi_2 \)-axiomatized theories. We denote it here as \( Z(\mathbf{EP}_2^c) \).

\[
Z(\mathbf{EP}_2^c) \quad (A \equiv_{\Pi_1} B) \rightarrow A \vdash_{\Pi_1} A \land B \quad \text{for} \ A \ \text{and} \ B \ \text{in} \ \mathbf{EP}_2^c
\]

where the class \( \mathbf{EP}_2^c \) of modal formulas is defined as follows:

\[
\mathbf{EP}_2^c := \Box A \mid \neg \Box A \mid \mathbf{EP}_2^c \land \mathbf{EP}_2^c \mid \mathbf{EP}_2^c \lor \mathbf{EP}_2^c \mid A \triangleright A.
\]

The class \( \mathbf{EP}_2^c \) is of course tailored so that any arithmetical realization will be provably \( \Pi_2 \). Note that the superscript c is there to indicate that the \( \triangleright \) modality is to be interpreted as a formalization of the notion of \( \Pi_1 \) conservativity. It is not hard to see that the formalization of this notion is itself \( \Pi_2 \). Moreover, note that this class coincides in extension with the earlier defined class \( \mathbf{ES}_2 \).

Since PRA is \( \Pi_2 \)-axiomatized and proves totality of the supexp function the principle \( Z(\mathbf{EP}_2^c) \) applies to PRA.

But there are repercussions for the interpretability logic of PRA as well. We know that for reflexive theories \( \Pi_1 \)-conservativity coincides with interpretability. We also know that any \( \Sigma_2 \)-extension of PRA is reflexive (**Lemma 1**). Altogether this means that a statement \( \alpha \vdash \beta \) and \( \alpha \vdash_{\Pi_1} \beta \) are equivalent if \( \alpha \) is in \( \Sigma_2 \) and PRA + \( \alpha \) is \( \Pi_2 \)-axiomatized, i.e. \( \alpha \) is in \( \Delta_2 \).
We arrive at Zambella’s principle for interpretability logic:

\[ Z (A \equiv B) \rightarrow A \triangleright A \land B \]  for \( A \) and \( B \) in \( BS_1 \)

For the \( \Pi_1 \)-conservativity logic of PRA, the principle \( Z(EP^\equiv) \) is really informative (see [3]), it is the only principle known on top of the basic ones for the \( \Pi_1 \)-conservativity logic of PRA. The principle \( Z \) for interpretability logic is very interesting as well but it does turn out to be derivable in \( ILB \) as we will now proceed to show. (See however the final remark of this section.)

Here modal logic again proves to be informative – to have such a proof is interesting since it is not at all clear to us how the two principles relate arithmetically.

We shall give a purely syntactical proof of \( ILB_0 \vdash Z \), \( B_0 \) being a restriction of \( B \) to \( BS_1 \) formulas, see Definition 5. The proof in [10] of the same fact was not correct.

Throughout the proof we consider a full disjunctive normal form of modal formulas:

**Definition 8.** A full disjunctive normal form (a full DNF) over a finite set of formulas \( \{C_1, \ldots, C_n\} \) is a disjunction of conjunctions of the form \( \pm C_1 \land \cdots \land \pm C_n \) where \( + C_i \) means \( C_i \) and \( - C_i \) means \( \neg C_i \), i.e., each \( C_i \) occurs either positively or negatively in each disjunct.

Each propositional formula is clearly equivalent to a formula in full DNF over the set of propositional atoms occurring in it. Similarly each modal \( BS_1 \)-formula, being a boolean combination of boxed formulas, is equivalent to a formula in full DNF over the set of its boxed subformulas, or even over any finite set of boxed formulas containing its boxed subformulas (or just its boxed subformulas maximal w.r.t. box-depth).

**Theorem 5.** \( ILB_0 \vdash Z \)

**Proof.** Let \( A, B \in BS_1 \) and let \( \{A_1, \ldots, A_m\} \) be the set of boxed subformulas of both \( A \) and \( B \). Assume w.l.o.g. that \( A \) and \( B \) are in full DNF over \( \{A_1, \ldots, A_m\} \). Assume \( A \equiv B \). We show that \( A \triangleright A \land B \). Since \( A \) comes in full DNF, this means to show, for each disjunct \( D \) of \( A \), that \( D \triangleright A \land B \). In fact, we show this for any disjunct of \( A \) or \( B \).

A disjunct \( D \) of either \( A \) or \( B \) is fully determined by the set \( D^\equiv \) of boxed formulas occurring positively in it. We shall write \( D^\equiv \) also for the conjunction of its members.

We first show, if \( D \) is a member of \( A \) or \( B \) which has a maximal set \( D^\equiv \) (no disjunct \( E \) with \( E^\equiv \) properly containing \( D^\equiv \) occurs in \( A \) or \( B \)) then \( D \triangleright A \land B \).

Suppose such \( D \) is in \( A \), the other case is symmetrical. Since \( D \triangleright A \) we have also \( D \triangleright B \). Then, noting that \( D^\equiv \) is a conjunction of boxed formulas and applying \( B_0 \), we obtain \( D \triangleright B \land D^\equiv \).

Now take any disjunct \( E \) of \( B \) for which \( E^\equiv \) does not contain \( D^\equiv \). Then \( E \) contradicts \( D^\equiv \) by its negative part. We distinguish two cases: if for all \( E \) in \( B \) the set \( E^\equiv \) does not contain \( D^\equiv \), then \( B \) contradicts \( D^\equiv \). It follows from \( D \triangleright B \land D^\equiv \) that \( D \not\triangleright \perp \). Then clearly \( D \triangleright A \land B \).

Otherwise \( B \) does contain \( E^\equiv \) with \( E^\equiv \) containing \( D^\equiv \). But since \( D \) has a maximal Box-set, \( E \) and \( D \) must be the same and \( D \) occurs in \( B \) as well. Thus \( D \triangleright B \land D \) and, since \( \vdash D \rightarrow A \), also \( D \triangleright A \land B \).

We have shown that all maximal disjuncts interpret \( A \land B \).

We show by induction that the same is true for all other disjuncts of \( A \) and \( B \). This suffices for the proof.

Assume that, for all \( k' \) with \( m \geq k' \) and all disjuncts \( D \) in either \( A \) or \( B \) with \( D^\equiv \) of size \( k' \), \( D \triangleright A \land B \) (this has already been shown for \( k \) equal to the size of the maximal Box-set in \( A \) and \( B \) which is certainly less then \( m \)). Consider a disjunct \( D \) of \( A \), the other case is again symmetrical. Assume w.l.o.g. that \( D^\equiv \) has size \( k \).

Since \( D \triangleright A \) and hence \( D \triangleright B \), we again have that \( D \triangleright B \land D^\equiv \). Now \( D^\equiv \) conflicts with all the disjuncts of \( B \), Box-set of which is not a superset of \( D^\equiv \). Again, we distinguish two cases: if there are no disjuncts of \( B \) with a Box-set which is a superset of \( D^\equiv \) then \( B \) conflicts with \( D^\equiv \) and \( D \not\triangleright \perp \) and thus \( D \triangleright A \land B \).

Otherwise some disjuncts of \( B \) do have a Box-set which is a superset of \( D^\equiv \). Let \( E_1, \ldots, E_i \) be all such disjuncts of \( B \). Then, since \( D \triangleright B \land D^\equiv \) and \( \vdash B \land D^\equiv \rightarrow E_1 \lor \cdots \lor E_i \) (where \( E_1 \lor \cdots \lor E_i \) is the part of \( B \) not conflicting with \( D^\equiv \)), we obtain \( D \triangleright E_1 \lor \cdots \lor E_i \). Now it suffices to show that each \( E_i \) interprets \( A \land B \).

Fix an \( E_i \) and suppose \( E_i^\equiv \) have size \( k \). But then \( E_i \equiv D \) and thus we have, as before, \( D \triangleright (B \land D) \) \( \rightarrow (B \land A) \). If \( E_i^\equiv \) have size greater then \( k \), the induction hypothesis apply and we obtain that \( E_i \) interprets \( A \land B \). □

Actually it is possible to extend Zambella’s principle somewhat in such a way that it is no longer clear whether the result is still derivable from \( B \). First note that the formulas in \( ES_1 \) are just the propositional combinations of \( \Box \)-formulas.

Zambella’s principle for interpretability logic as studied in this paper reads

\[ A \equiv B \rightarrow A \triangleright A \land B \]

where \( A \) and \( B \) should both be \( BS_1 \). However, to have access to the ideas behind Zambella’s principle, it is sufficient that \( A \) and \( B \) be both provably of complexity \( \Delta_2 \). We can thus look at those \( ES_2 \) formulae which are provably equivalent to the negation of some other \( ES_2 \) formula and plug those formulae in. Reflecting this thought in a formula yields\(^3\)

\[ \Box((A \leftrightarrow A') \land (B \leftrightarrow B')) \rightarrow (A \equiv B \rightarrow A \triangleright A \land B) \]

\(^3\) We would like to thank one of the referees for pointing out that our original extension of Zambella’s principle for interpretability logic could actually be even generalized to its current form.
where \(A, A', B\) and \(B'\) are all from \(ES_2\). It actually makes sense to call this principle the Zambella principle for interpretability logic as it more precisely reflects the arithmetical ingredients. We have chosen not to do so as to be consistent with earlier papers.

8. Delimitation of IL(PRA)

Let us see what we can conclude about IL(PRA) from the above. Certainly IL(PRA) includes IL(All) but it is more than that because \(B\) is not a principle of IL(All). The latter is clear from the fact that IL(All) \(\subseteq ILM \cap ILP\) and \(Z\) is not in ILP: consider the following model:

\[
\begin{align*}
& A \models \diamond p \equiv \diamond q \\
& w \not\models p \land q, \\
& \text{thus Zambella fails.}
\end{align*}
\]

This shows, by derivability of \(Z\) from \(B\), that indeed \(B\) is not a principle of IL(All).

Also we know that IL(PRA) is not ILM since \(M\) is not in IL(PRA), as A. Visser discusses in [17]: the two logics cannot be the same because if ILM is a part of the interpretability logic of a theory then it is a part of the interpretability logic of any of its finite extensions as well. This cannot be the case for PRA because not all of its finite extensions are reflexive. A more specific example of a principle of ILM which is not in IL(PRA) can be given:

\[
A \subseteq \diamond B \rightarrow \Box(A \subseteq \diamond B).
\]

That this formula is not in IL(PRA) can be shown using Shavrukov’s result from [14] about complexity of the set \(\{\psi \mid \psi \in \Pi_1 \land \phi \supset \psi\}\); see [17] for the full proof.

We know that \(M_0\) is provable in ILB. The other principles surely contained in IL(PRA) are \(B, R\) and \(W\) (\(R^*\) is the conjunction of \(R\) and \(W\)). Let us show they are mutually independent. Note that for nonderivability proofs soundness suffices.

First let us recall the frame conditions for the two principles \(W\) and \(R\). The condition for \(W\) requires that the composition \((S_w; R)\) is inversely well-founded, the condition for \(R\) is the following: \(xRyRzS_xuRv \Rightarrow zSyv\).

\(W\) vs. \(B\): It is easy to see that \(W \not\vdash B\) since the former is in IL(All) while the later is not in it. Since \(R\) is in IL(All) as well, \(W, R \not\vdash B\). The following frame

\[
\begin{align*}
& wRyRzS_xuRv \\
& wRz \text{ is bi-similar to } y \text{ and } B \text{ is ensured. Thus } B \not\vdash W.
\end{align*}
\]

Moreover, the same frame, being an \(R\) frame, shows that \(B, R \not\vdash W\): the only case to check is \(wRyRzS_xxRy\). Now the condition for \(R\) requires \(yS_yx\), but this is clearly the case since \(S_y\) is reflexive over \(x\).

\(R\) vs. \(B\): Again, since \(R \in IL(All)\), it cannot be that \(R \vdash B\). We have already discussed that neither \(R, W \vdash B\). The following frame

\[
\begin{align*}
& wRyRzS_xuRv \\
& wRz \text{ has no successors at all.}
\end{align*}
\]

Moreover, since the frame is clearly a \(W\) frame as well, we have shown that \(B, W \not\vdash R\).

\(R\) vs. \(W\): already discussed in [8].
It is clear from our exposition that, although we have solved a number of problems concerning $\mathbf{IL}(\text{PRA})$, many remain open, e.g. those connected with our incomplete knowledge of $\mathbf{IL}(\text{All})$. Also, we lack a modal completeness theorem for $\mathbf{IL}_B$. Unfortunately, the complexity of the frame condition for $B$ makes this seem an intractable problem at the present time. In any case, the logic of interpretability is far from being a finished subject.

Acknowledgment

First author was supported by grants GA CR 401/06/0387 and IAA900090703.

References

[1] L. Beklemishev, Induction rules, reflection principles, and provably recursive functions, Annals of Pure and Applied Logic 85 (1997) 193–242.
[2] L. Beklemishev, Proof-theoretic analysis by iterated reflection, Archive for Mathematical Logic 42 (6) (2003) 515–552.
[3] L. Beklemishev, A. Visser, On the limit existence principles in elementary arithmetic and related topics, in: Logic Group Preprint Series, vol. 224, University of Utrecht, 2004.
[4] A. Berarducci, The interpretability logic of Peano arithmetic, Journal of Symbolic Logic 55 (1990) 1059–1089.
[5] D. de Jongh, G. Japaridze, in: S. Buss (Ed.), The Logic of Provability, in: Handbook of Proof Theory. Studies in Logic and the Foundations of Mathematics, vol. 137, Elsevier, Amsterdam, 1998, pp. 475–546.
[6] D. de Jongh, F. Veltman, Provability logics for relative interpretability, in: [12], 1990.
[7] D. de Jongh, F. Veltman, Modal completeness of $\mathbf{IL}_W$, in: J. Gerbrandy, M. Marx, M. Rijke, Y. Venema (Eds.), Essays Dedicated to Johan van Benthem on the Occasion of His 50th Birthday, Amsterdam University Press, Amsterdam, 1999.
[8] E. Goris, J. Joosten, Modal matters in interpretability logics, in: Logic Group Preprint Series, vol. 226, University of Utrecht, March 2004.
[9] P. Hájek, F. Montagna, The logic of $\Pi^1_1$-conservativity, Archiv für Mathematische Logik und Grundlagenforschung 30 (1990) 113–123.
[10] J.J. Joosten, Interpretability formalized, Ph.D. Thesis, Utrecht University, November 2004. ISBN: 90-393-3869-8.
[11] J.J. Joosten, E. Goris, A new principle in the interpretability logic of all reasonable arithmetical theories, The Logic Journal of the IGPL (in press).
[12] P. Petkov (Ed.), Mathematical logic, in: Proceedings of the Heyting 1988 Summer School in Varna, Bulgaria, Plenum Press, Boston, 1990.
[13] V. Shavrukov, The logic of relative interpretability over Peano arithmetic, Tech. Rep. Report No.5, Steklov Mathematical Institute, Moscow, 1988 (in Russian).
[14] V. Shavrukov, Interpreting reflexive theories in finitely many axioms, Fundamenta Mathematicae 152 (1997) 99–116.
[15] A. Tarski, A. Mostowski, R. Robinson, Undecidable theories, North Holland, Amsterdam, 1953.
[16] A. Visser, Interpretability logic, in: [12], 1990.
[17] A. Visser, An overview of interpretability logic, in: M. Kracht, M.D. Rijke, H. Wansing (Eds.), Advances in Modal Logic '96, CSLI Publications, Stanford, CA, 1997, pp. 307–359.
[18] D. Zambella, Chapters on bounded arithmetic & on provability logic, Ph.D. Thesis, University of Amsterdam, 1994.