Interpretability in PRA *

Marta Bílková Dick de Jongh Joost J. Joosten

June 30, 2008

Abstract

In this paper we study **IL**(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: **IL**(PRA) is not **IL**M or **IL**P. **IL**(PRA) does of course contain all the principles known to be part of **IL**(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 **IL**(PRA) are. These properties are reflected in the so-called Beklemishev Principle B, and Zambella's Principle Z, neither of which is a part of **IL**(All). Both principles and their interrelation are submitted to a modal study. In particular, we prove a frame condition for B. morover, we prove that Z follows from a restricted form of B. Finally, we give an overview of the known relationships of **IL**(PRA) to important other interpetability 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 $T \rhd S$, which then, roughly, means that there is a translation $\cdot^t$ 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, $(\varphi \vee \psi)^t = \varphi^t \vee \psi^t$ and in particular $\bot^t = (\vee_\emptyset)^t = \vee_\emptyset = \bot$. We refer the reader to [17], [5] and [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:

$$(T + \alpha \rhd T + \beta) \wedge (T + \beta \rhd T + \gamma) \rightarrow (T + \alpha \rhd T + \gamma).$$

*dedicated to Franco Montagna on his sixtieth birthday

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.

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 $\mathbf{IL}\mathsf{M}$. The principle $(A \rhd B) \rightarrow (A \wedge \Box C \rhd B \wedge \Box C)$ which 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 interpetability 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 $\mathbf{IL}\mathsf{P}$. 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 - $\mathbf{IL}(\mathrm{All})$. As PRA is certainly a 'reasonable arithmetical theory', this core logic should also be a part of $\mathbf{IL}(\mathrm{PRA})$. 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 principal 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 Σ_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 Π_1-conservativity logics. A theory S is Π_1 conservative over a theory T in the same language of arithmetic, we shall write $S \rhd_{\Pi_1} T$ whenever S proves any Π_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 Σ_1 sentence σ, the following is a valid principle $S \rhd_{\Pi_1} T \rightarrow S + \sigma \rhd_{\Pi_1} T + \sigma$. 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 $S \rhd T \leftrightarrow S \rhd_{\Pi_1} T$. This equivalence was exploited by Hájek and Montagna who were the first to show that the Π_1-conservativity logic of PA is $\mathbf{IL}\mathsf{M}$ as well

[10]. The observation about the equivalnce is more generally important when looking at the repercussions of Π_1-conservativity principles on interpretability logics. In this paper we shall consider Zambella's principle for Π_1-conservativity logics and look at its repercussions for the interpretability logic of PRA. We shall show that Zambella does not add new information in the sense that its modal-logical consequnces are already implied by Beklemishev's principle.

It is remarkable that the notion of interpretability is, in a sense, less stable than that of Π_1-conservativity. Hájek and Montagna show that their results extends to all reasonable theories containing $\mathrm{I}\Sigma_1$. This was strengthened by Beklemishev and Visser in [3]: all theories extending the parameter-free induction schema $\mathrm{I}\Pi_1^-$ have the same Π_1-conservativity logic (**IL**M) whereas in this range the interpretability logics expose a diverse and wild behavior. Note though that PRA does not prove $\mathrm{I}\Pi_1^-$, and, in fact, the Π_1-conservativity logic of PRA remains unknown.

A number of the results in this paper was first proved in [11].

2 Arithmetic

Let us first fix some arithmetical notation. We use modal symbols $\Box, \Diamond, \rhd$ both in modal and arithmetical statements, here we fix their arithmetical meaning. We write, for an arithmetical sentence α, $\Box_{\mathrm{T}}\alpha$ for formalized provability in T, $\Box_{\mathrm{T},n}\alpha$ for formalized provability of α in T using only non-logical axioms with Gödel numbers $\leq n$ and formulas of logical complexity $\leq n$. Dually, $\Diamond_{\mathrm{T}}\alpha = \neg\Box_{\mathrm{T}}\neg\alpha$ means formalized consistency of α over T (i.e. nonexistence of a proof of a contradiction from α), while $\Diamond_{\mathrm{T},n}\alpha$ means $\neg\Box_{\mathrm{T},n}\neg\alpha$. For theories T, S we use $\mathrm{T} \rhd \mathrm{S}$ to denote formalized interpretability of S in T. For arithmetical sentences α, β, $\alpha \rhd_{\mathrm{T}} \beta$ means $\mathrm{T} + \alpha \rhd \mathrm{T} + \beta$. Similarly for theories T, S, $\rhd_{\Pi_1}$ denotes formalized Π_1-conservativity of T over S and for arithmetical sentences α, β, $\alpha \rhd_{\Pi_1} \beta$ means $\mathrm{T} + \alpha \rhd_{\Pi_1} \mathrm{T} + \beta$.

2.1 What is PRA?

In the literature there are many definitions of PRA around. 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 Δ_0-formula in this enriched language.

Beklemishev has shown in [2] that PRA is in a strong sense equivalent (faithfully bi-interpretable) with $(\mathrm{EA})_2^\omega$. Here, $(\mathrm{EA})_2^\omega$ is the theory that is obtained by starting with EA ($= \mathrm{I}\Delta_0 + \mathsf{exp}$) and iterating '$\omega$ many times' Π_2-reflection. In symbols: $(\mathrm{EA})_2^0 = \mathrm{EA}$, and $(\mathrm{EA})_2^{n+1} = \mathsf{RFN}_{(\mathrm{EA})_2^n}(\Pi_2)$.

In this paper, we shall use the definition:

$$\mathrm{PRA} := (\mathrm{EA})_2^\omega.$$

Under this definition, the following lemma is immediate.

Lemma 2.1. *Any r.e. extension of* PRA *by* Σ_2^0 *sentences is reflexive.*

2.2 The Orey-Hájek Characterizations for interpretability

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 2.2. *Whenever T is reflexive we have that*

$$T \rhd S \quad \Leftrightarrow \forall x \; T \vdash \neg \Box_{S,x} \bot$$

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

$$\vdash T \rhd S \leftrightarrow \forall x \; \Box_T \neg \Box_{S,x} \bot$$

In [11] an overview is given of all the implications, corresponding requirements and necessary arguments regarding Orey-Hájek. In the above Lemma the $\Leftarrow$ 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 $\mathtt{exp}$ as well.

Note that, using the above characterization, the a-priori Σ_3 notion of interpretability becomes Π_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 fenomena since it can visualize their behaviour using a simple language and an intuitive frame semantics. Perhaps the most significant point where modal logic shows its skills are completeness proofs - arithmatical 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, 11, 8].

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

$$\mathcal{A} ::= p | \bot | \mathcal{A} \wedge \mathcal{A} | \mathcal{A} \to \mathcal{A} | \Box \mathcal{A} | \mathcal{A} \rhd \mathcal{A}$$

We will use standard abbreviations $\Diamond, \vee, \neg, \top, \leftrightarrow$, and we write $A \equiv B$ instead of $(A \rhd B) \wedge (B \rhd A)$.

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 $*$ commutes with all boolean connectives. Moreover $(\Box A)^* = \Box_T A^*$ and $(A \rhd B)^* = A^* \rhd_T B^*$, 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 A such that $\forall * \, T \vdash A^*$. The *interpretability logic* of a theory T, denoted **IL**(T), is then the set of all the interpretability principles of T.

3.1 The logic IL

The logic **IL** is in a sense the core interpretability logic - it is a (proper) part of the interpretability logic of any reasonable arithmetical theory: $\mathbf{IL} \subset \mathbf{IL}(\mathrm{T})$. It captures the basic structural behaviour of interpretability.

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:

L1	$\Box(A \to B) \to (\Box A \to \Box B)$
L2	$\Box A \to \Box\Box A$
L3	$\Box(\Box A \to A) \to \Box A$
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	$\Diamond A \rhd A$

Note that the part of **IL** not containing the $\rhd$ modality is the well-known Gödel-Löb provability logic **GL**, axiomatized by the first three schemata. It is easy to show that $\Box$ can be defined in terms of $\rhd$ modality: $\vdash_{\mathbf{IL}} \Box A \leftrightarrow \neg A \rhd \bot$.

More interpretability logics are obtained extending **IL** by new interpretability principles. Some of such principles are listed below:

W	$A \rhd B \to A \rhd B \wedge \Box\neg A$
W*	$A \rhd B \to B \wedge \Box C \rhd B \wedge \Box C \wedge \Box\neg A$
M$_0$	$A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C$
M	$A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$
P	$A \rhd B \to \Box(A \rhd B)$
R	$A \rhd B \to \neg(A \rhd \neg C) \rhd B \wedge \Box C$
R*	$A \rhd B \to \neg(A \rhd \neg C) \rhd B \wedge \Box C \wedge \Box\neg A$

All of these principles are in **IL**(All) except the principles M and P which were mentioned above already. For an overview, see [17] and [8]. For the last word on **IL**(All) see [9].

For X a set of principles we denote **IL**X the logic extending **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 **IL**M. For finitely axiomatizable theories containing supexp the interpretability logic is known to be **IL**P ([16]).

An important consequence of **IL**M that expresses the Π_1-conservativity of interpretability more directly is $(A \rhd \Diamond B) \to \Box(A \to \Diamond B)$.

3.2 Modal semantics

Modal frame semantics of interpretability logics is based on **GL**-frames extended with a ternary accesibility relation interpreting the binary $\rhd$ modality. The ternary relation is however given by a set of binary relations indexed by the nodes:

Definition 3.1. An **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. $yS_xz \to xRy \& xRz$
3. $xRy \to yS_xy$
4. $xRyRz \to yS_xz$
5. $uS_xvS_xw \to uS_xw$

An **IL**-model is a quadruple $\langle W, R, S, \Vdash\rangle$ where $\langle W, R, S\rangle$ is a **IL**-frame and $\Vdash$ is a subset of $W \times \mathsf{Prop}$, extending to boolean formulas as usualy and to modal formulas as follows:

$$
\begin{array}{lll}
w & \Vdash & \Box A \text{ iff } \forall v(wRv \Rightarrow v \Vdash A) \\
w & \Vdash & A \rhd B \text{ iff } \forall u(wRu \;\&\; u \Vdash A \Rightarrow \exists v(uS_wv \Vdash B))
\end{array}
$$

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 $\mathcal{C}$ in first or higher order logic is a *frame condition* for X if, for each frame F,

$$F \models \mathcal{C} \text{ iff } F \models \mathsf{X}.$$

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

M	$xRyS_xzRu \Rightarrow yRu$
$\mathsf{M_0}$	$xRyRzS_xuRv \Rightarrow yRv$
P	$xRyRzS_xu \Rightarrow yRu \wedge zS_yu$
W	$(S_w; R)$ is conversely well-founded
R	$xRyRzS_xuRv \Rightarrow zS_yv$

We have the following completeness results: **IL** is sound and complete w.r.t. (finite) **IL** frames, **IL**P is complete w.r.t. (finite) **IL**P frames (all in [6]), **IL**W is complete w.r.t. (finite) **IL**W frames ([7], see also [8]), **IL**M is complete w.r.t. (finite) **IL**M 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 Σ_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 Σ_2 in PRA. These are called *essentially* Σ_2*-formulas*, we write ES_2. Let us start by defining this class and some related classes. In our definition, $\mathcal{A}$ will stand for the set of all modal interpretability formulae.

$$\begin{array}{lcl}
\mathsf{ED}_2 & := & \Box\mathcal{A} \mid \neg\mathsf{ED}_2 \mid \mathsf{ED}_2 \wedge \mathsf{ED}_2 \mid \mathsf{ED}_2 \vee \mathsf{ED}_2 \\
\mathsf{ES}_2 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathsf{ES}_2 \wedge \mathsf{ES}_2 \mid \mathsf{ES}_2 \vee \mathsf{ES}_2 \mid \neg(\mathsf{ES}_2 \rhd \mathcal{A}) \\
\mathsf{EP}_2^{\mathsf{c}} & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathsf{EP}_2^{\mathsf{c}} \wedge \mathsf{EP}_2^{\mathsf{c}} \mid \mathsf{EP}_2^{\mathsf{c}} \vee \mathsf{EP}_2^{\mathsf{c}} \mid \mathcal{A} \rhd \mathcal{A} \\
\mathsf{ES}_3 & := & \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathsf{ES}_3 \wedge \mathsf{ES}_3 \mid \mathsf{ES}_3 \vee \mathsf{ES}_3 \mid \mathcal{A} \rhd \mathcal{A} \\
\mathsf{ES}_4 & := & \Box\mathcal{A} \mid \neg\mathsf{ES}_4 \mid \mathsf{ES}_4 \wedge \mathsf{ES}_4 \mid \mathsf{ES}_4 \vee \mathsf{ES}_4 \mid \mathcal{A} \rhd \mathcal{A}
\end{array}$$

We can now formulate Beklemishev's principle B.

$$\mathsf{B} := A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C \qquad \text{for } A \in \mathsf{ES}_2$$

Note that B is just Montagna's principle M restricted to ES_2-formulas.

Lemma 4.1. $\mathbf{IL}\mathsf{B} \vdash \mathsf{B}'$*, where* $\mathsf{B}' : \mathsf{A} \rhd \mathsf{B} \to \mathsf{A} \wedge \mathsf{C} \rhd \mathsf{B} \wedge \mathsf{C}$ *with* $A \in \mathsf{ES}_2$ *and* C *a CNF (a conjunction of disjunctions) of boxed formulas.*

Proof. Easy. □

5 Arithmetical soundness of B

By Lemma 2.1 we know that $\mathrm{PRA} + \sigma$ is reflexive for any $\Sigma_2(\mathrm{PRA})$-sentence σ. Thus, we get by Orey-Hájek that

$$\mathrm{PRA} \vdash \sigma \rhd_{\mathrm{PRA}} \psi \leftrightarrow \forall x\ \Box_{\mathrm{PRA}}(\sigma \to \Diamond_{\mathrm{PRA},x}\psi). \tag{1}$$

Consequently, for $\sigma \in \Sigma_2(\mathrm{PRA})$, $\neg(\sigma \rhd_{\mathrm{PRA}} \psi) \in \Sigma_2(\mathrm{PRA})$ and we see that, indeed, $\forall\, A{\in}\mathsf{ES}_2\, \forall *\ \ A^* \in \Sigma_2(\mathrm{PRA})$. This enables us to prove the arithmetical soundness of B.

Theorem 5.1. *For any formulas* B *and* C *we have that* $\forall\, A{\in}\mathsf{ES}_2\, \forall *$ $\mathrm{PRA} \vdash (A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C)^*$.

Proof. For some $A \in \mathsf{ES}_2$ and arbitrary B and C, we consider some realization $*$ and let $\alpha := A^*$, $\beta := B^*$ and $\gamma := C^*$. We reason in PRA and assume $\alpha \rhd_{\mathrm{PRA}} \beta$. As α is $\Sigma_2(\mathrm{PRA})$, we get by (1) that

$$\forall x\ \Box_{\mathrm{PRA}}(\alpha \to \Diamond_{\mathrm{PRA},x}\beta). \tag{2}$$

We now consider n large enough (dependent on γ) such that

$$\Box_{\mathrm{PRA}}(\Box_{\mathrm{PRA}}\gamma \to \Box_{\mathrm{PRA},n}\Box_{\mathrm{PRA}}\gamma). \tag{3}$$

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

$$\Box_{\mathrm{PRA},n}(\delta \to \neg\epsilon) \wedge \Box_{\mathrm{PRA},n}\delta \to \Box_{\mathrm{PRA},n}\neg\epsilon,$$

whence

$$\Diamond_{\mathrm{PRA},n}\epsilon \wedge \Box_{\mathrm{PRA},n}\delta \to \Diamond_{\mathrm{PRA},n}(\delta \wedge \epsilon) \tag{4}$$

Combining (2), (3), and using (4), we see that for any x, $\Box(\alpha \wedge \Box\gamma \to \Diamond_{\mathrm{PRA},n}(\beta \wedge \Box\gamma))$. Clearly, $\alpha \wedge \Box\gamma$ is still a $\Sigma_2(\mathrm{PRA})$-sentence.[1] Again by (1) we get $\alpha \wedge \Box\gamma \rhd \beta \wedge \Box\gamma$. □

[1] Actually, this observation is not necessary as we use the direction in the Orey-Hájek Characterization that does not rely on the reflexivity.

Let $\mathsf{M}^{\mathsf{ES_n}}$ be the schema $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$ with $A \in \mathsf{ES_n}$. Theorem 5.1 can be generalized using results of [1] to the theory $\mathrm{I}\Sigma_n^R$, which is Robinson's arithmetic Q plus the Σ_n induction *rule*, for $n = 1, 2, 3$ as follows:

Theorem 5.2. $\mathbf{IL}(\mathrm{I}\Sigma_{\mathrm{n}}^{\mathrm{R}}) \vdash \mathsf{M}^{\mathsf{ES_{n+1}}}$ *for* $n = 1, 2, 3$.

6 A frame condition for B

Let us first fix some notation. If $\mathcal{C}$ is a finite set, we write $xR\mathcal{C}$ as short for $\bigwedge\!\!\!\bigwedge_{c\in\mathcal{C}} xRc$. Similar conventions hold for the other relations. The A-critical cone of x, $\mathcal{C}_x^A$ is in this section defined as $\mathcal{C}_x^A := \{y \mid xRy \wedge \forall z\ (yS_xz \to z \nVdash A)\}$.

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

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 $\Vdash$ 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 6.1.

$$\mathsf{ES}_2^0 \ := \ \mathsf{ED}_2$$
$$\mathsf{ES}_2^{n+1} \ := \ \mathsf{ES}_2^{n} \mid \mathsf{ES}_2^{n+1} \wedge \mathsf{ES}_2^{n+1} \mid \mathsf{ES}_2^{n+1} \vee \mathsf{ES}_2^{n+1} \mid \neg(\mathsf{ES}_2^{n} \rhd \mathcal{A})$$

It is clear that $\mathsf{ES}_2 = \cup_i \mathsf{ES}_2^i$. We now define some first order formulas $\mathcal{S}_i(b, u)$ that say that two nodes in a frame b and u look alike. The larger i is, the more the two points look alike. We use the letter $\mathcal{S}$ as to hint at a simulation.

Definition 6.2.

$$\mathcal{S}_0(b, u) \ := \ b{\uparrow}{=}u{\uparrow}$$
$$\mathcal{S}_{n+1}(b, u) \ := \ \mathcal{S}_n(b, u) \wedge \forall c\ (bRc \to \exists c'\ (uRc' \wedge \mathcal{S}_n(c, c') \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow}))$$

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

Lemma 6.3. *Let F be a model. For all n we have the following. If $F \models \mathcal{S}_n(b, u)$, then $b \Vdash A \Rightarrow u \Vdash A$ for all $A{\in}\mathsf{ES}_2^n$.*

Proof. We proceed by induction on n. If $n{=}0$, $A{\in}\mathsf{ES}_2^0$ can be written as $\bigvee\!\!\!\bigvee_i (\Box A_i \wedge \bigwedge\!\!\!\bigwedge_j \Diamond A_{ij})$. Clearly, if $b{\uparrow} = u{\uparrow}$ then $b \Vdash A \Rightarrow u \Vdash A$.

Now consider $A{\in}\mathsf{ES}_2^{n+1}$ and b and u such that $F \models \mathcal{S}_{n+1}(b, u)$. We can write

$$A = \bigvee\!\!\!\bigvee_i (A_{i0} \wedge \bigwedge\!\!\!\bigwedge_{j\neq 0} \neg(A_{ij} \rhd B_{ij})),$$

with A_{ij} in $\mathsf{ES}_2^{\mathsf{n}}$. If $b \Vdash A$, then for some i, $b \Vdash A_{i0} \wedge \bigwedge\!\!\!\bigwedge_{j \neq 0} \neg(A_{ij} \rhd B_{ij})$. As $\mathcal{S}_{n+1}(b,u) \to \mathcal{S}_n(b,u)$, and by the induction hypothesis we see that $u \Vdash A_{i0}$. So, we only need to see that $u \Vdash \neg(A_{ij} \rhd B_{ij})$ for $j{\neq}0$. As $b \Vdash \neg(A_{ij} \rhd B_{ij})$, for some $c \in \mathcal{C}_b^{B_{ij}}$ we have $c \Vdash A_{ij}$. By $\mathcal{S}_{n+1}(b,u)$ we find a c' such that uRc', and $c'S_u{\uparrow} \subseteq cS_b{\uparrow}$ (thus cS_bc'). This guarantees that $c' {\in} \mathcal{C}_u^{B_{ij}}$. Moreover we know that $\mathcal{S}_n(c,c')$, thus by the induction hypothesis, as $c \Vdash A_{ij}$, we get that $c' \Vdash A_{ij}$. Consequently $u \Vdash \neg(A_{ij} \rhd B_{ij})$.

□

Lemma 6.4. *Let F be a finite frame. For all i, and any $b{\in}F$, there is a valuation V_i^b on F and a formula $A_i^b {\in} \mathsf{ES}_2^{\mathsf{i}}$ such that $F \models \mathcal{S}_i(b,u) \Leftrightarrow u \Vdash A_i^b$.*

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 $\{x_j\}_{j \in J}$. We define

$$\begin{array}{lcl} y \Vdash p_j & \leftrightarrow & y{=}x_j \\ y \Vdash r & \leftrightarrow & bRy. \end{array}$$

Let A_0^b be $\Box r \wedge \bigwedge\!\!\!\bigwedge_j \Diamond p_j$. It is now obvious that $u \Vdash A_0 \Leftrightarrow u{\uparrow}{=}b{\uparrow}$.

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

Let $\{q_j\}_{j \in J}$ be a set of fresh variables. V_{i+1}^b will be V_i^b and V_i^j on the old variables. For the $\{q_j\}_{j \in J}$ we define V_{i+1}^b to act as follows:

$$y \Vdash q_j \Leftrightarrow y \notin x_j S_b{\uparrow}.$$

Moreover we define

$$A_{i+1}^b := A_i^b \wedge \bigwedge\!\!\!\bigwedge_j \neg(A_i^j \rhd q_j).$$

Now we will see that under the new valuation V_{i+1}^b,

(*i*) $u \Vdash A_{i+1}^b \Rightarrow F \models \mathcal{S}_{i+1}(b,u)$,

(*ii*) $F \models \mathcal{S}_{i+1}(b,u) \Rightarrow u \Vdash A_{i+1}^b$.

For (*i*) we reason as follows. Suppose $u \Vdash A_{i+1}^b$. Then also $u \Vdash A_i^b$ and thus $F \models \mathcal{S}_i(b,u)$. It remains to show that

$$F \models \forall c\, (bRc \to \exists c'\, (uRc' \wedge \mathcal{S}_i(c,c') \wedge cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})).$$

To this purpose we consider and fix some x_j in $b{\uparrow}$. As $u \Vdash A_{i+1}^b$, we get that $u \Vdash \neg(A_i^j \rhd q_j)$. Thus, for some $c' \in \mathcal{C}_u^{q_j}$, $c' \Vdash A_i^j$. Clearly $c' \Vdash \neg q_j$ whence x_jS_bc'. Also $\forall t\ (c'S_uy \Rightarrow y \Vdash \neg q_j)$ which, by the definition of V_{i+1}^b translates to $c'S_u{\uparrow} \subseteq x_jS_b{\uparrow}$. Clearly also uRc'. By $c' \Vdash A_i^j$ and the induction hypothesis we get that $\mathcal{S}_i(x_j, c')$. Indeed we see that $F \models \mathcal{S}_{i+1}(b,u)$.

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

$$uRc' \wedge \mathcal{S}_i(x_j, c') \wedge x_j S_b c' \wedge c' S_u{\upharpoonright} \subseteq x_j S_b{\upharpoonright}.$$

Now, $uRc' \wedge x_j S_b c' \wedge c' S_u{\upharpoonright} \subseteq x_j S_b{\upharpoonright}$ gives us that $c' \in \mathcal{C}_u^{q_j}$. By $\mathcal{S}_i(x_j, c')$ and the induction hypothesis we get that $c' \Vdash A_i^j$. Thus indeed $u \Vdash \neg(A_i^j \rhd q_j)$. $\Box$

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

Definition 6.5. For every i we define the frame condition $\mathcal{C}_i$ to be

$$\forall\, a, b\; (aRb \to \exists u\; (bS_a u \wedge \mathcal{S}_i(b,u) \wedge \forall\, d, e\; (uS_a dRe \to bRe))).$$

Lemma 6.6. *Let F be a finite frame. For all i, we have that*

$$\text{for all } A{\in}\mathsf{ES}_2^{\mathsf{i}},\ F \models A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C,$$
if and only if
$$F \models \mathcal{C}_i.$$

Proof. First suppose that $F \models \mathcal{C}_i$ and that $a \Vdash A \rhd B$ for some $A{\in}\mathsf{ES}_2^{\mathsf{i}}$ and some valuation on F. We will show that $a \Vdash A \wedge \Box C \rhd B \wedge \Box C$ for any C. Consider therefore some b with aRb and $b \Vdash A \wedge \Box C$. The $\mathcal{C}_i$ condition provides us with a u such that

$$bS_a u \wedge \mathcal{S}_i(b,u) \wedge \forall\, d, e\; (uS_a dRe \to bRe) \quad (*)$$

As $F \models \mathcal{S}_i(b,u)$, we get by Lemma 6.3 that $u \Vdash A$. Thus, as aRu and $a \Vdash A \rhd B$, we know that there is some d with $uS_a d$ and $d \Vdash B$. If now dRe, by $(*)$, also bRe and hence $e \Vdash C$. Thus, $d \Vdash B \wedge \Box C$. Clearly $bS_a d$ and thus $a \Vdash A \wedge \Box C \rhd B \wedge \Box C$.

For the opposite direction we reason as follows. Suppose that $F \not\models \mathcal{C}_i$. Thus, we can find a, b with

$$aRb \wedge \forall u\; (bS_a u \wedge \mathcal{S}_i(b,u) \to \exists\, d, e\; (uS_a dRe \wedge \neg bRe)) \quad (**).$$

By Lemma 6.4 we can find a valuation V_i^b and a sentence $A_i^b{\in}\mathsf{ES}_2^{\mathsf{i}}$ such that $u \Vdash A_i^b \Leftrightarrow F \models \mathcal{S}_i(b,u)$. Let q and s be fresh variables. Moreover, let $\mathcal{D}$ be the following set.

$$\mathcal{D} := \{d{\in}F \mid bS_a dRe \wedge \neg bRe \text{ for some } e\ \}.$$

We define a valuation V that is an extension of V_i^b by stipulating that

$$\begin{array}{lcl} y \Vdash q & \leftrightarrow & (y{\in}\mathcal{D}) \vee \neg(bS_a y), \\ y \Vdash s & \leftrightarrow & bRy. \end{array}$$

We now see that

(i) $a \Vdash A_i^b \rhd q$,

(ii) $a \Vdash \neg(A_i^b \wedge \Box s \rhd q \wedge \Box s)$.

For (i) we reason as follows. Suppose that aRb' and $b' \Vdash A_i^b$. If $\neg(bS_ab')$, $b' \Vdash q$ and we are done. So, we consider the case in which bS_ab'. As $\mathcal{S}_i(b, b')$, $(**)$ now yields us a $d \in \mathcal{D}$ such that $b'S_ad$. Clearly bS_ad and thus, by definition, $d \Vdash q$.

To see (ii) we notice that $b \Vdash A_i^b \wedge \Box s$. But if bS_ay and $y \Vdash q$, by definition $y \in \mathcal{D}$ and thus $y \Vdash \neg\Box s$. Thus $b \in \mathcal{C}_a^{q \wedge \Box s}$ and $a \Vdash \neg(A_i \wedge \Box s \rhd q \wedge \Box s)$. □

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

Theorem 6.7. *A finite frame F validates all instances of Beklemishev's principle if and only if $\forall i\ F \models \mathcal{C}_i$.*

Definition 6.8. Let $\mathsf{B_i}$ be the principle $A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C$ for $A \in \mathsf{ES}_2^{\mathsf{i}}$.

Corollary 6.9. *For a finite frame we have $F \models \mathsf{B_i} \Leftrightarrow F \models \mathcal{C}_i$.*

For the class of finite frames, we can get rid of the universal quantification in the frame condition of Beklemishev's principle. Remember that $\mathsf{depth}(x)$, the depth of a point x, is the length of the longest chain of R-successors starting in x.

Lemma 6.10. *If $\mathcal{S}_n(x, x')$, then $\mathsf{depth}(x) = \mathsf{depth}(x')$.*

Proof. $\mathcal{S}_n(x, x') \Rightarrow \mathcal{S}_0(x, x') \Rightarrow x{\uparrow} = x'{\uparrow}$. □

Lemma 6.11. *If $\mathcal{S}_n(x, x')$ & $\mathsf{depth}(x) \leq n$, then $\mathcal{S}_m(x, x')$ 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 $\mathcal{S}_{n+1}(x, x')$ & $\mathsf{depth}(x) \leq n + 1$. We are done if we can show $\mathcal{S}_{m+1}(x, x')$ for $m \geq n + 1$.

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

$$\mathcal{S}_m(x, x') \wedge \forall y\ (xRy \to \exists y'\ (yS_xy' \wedge \mathcal{S}_m(y, y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow}))$$

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

$$yS_xy' \wedge \mathcal{S}_n(y, y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow}.$$

As xRy and $\mathsf{depth}(x) \leq n + 1$, we see $\mathsf{depth}(y) \leq n$. Hence by the main induction, we get that $\mathcal{S}_m(y, y')$ and we are done. □

Definition 6.12. A B-simulation on a frame is a binary relation $\mathcal{S}$ for which the following holds.

1. $\mathcal{S}(x, x') \to x{\uparrow} = x'{\uparrow}$
2. $\mathcal{S}(x, x')$ & $xRy \to \exists y'(yS_xy' \wedge \mathcal{S}(y, y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow})$

If F is a finite frame that satisfies $\mathcal{C}_i$ for all i, we can consider $\bigcap_{i \in \omega} \mathcal{S}_i$. This will certainly be a B-simulation.

Definition 6.13. The frame condition $\mathcal{C}_{\mathsf{B}}$ is defined as follows. $F \models \mathcal{C}_{\mathsf{B}}$ if and only if there is a B-simulation $\mathcal{S}$ on F such that for all x and y,

$$xRy \to \exists y' (yS_x y' \wedge \mathcal{S}(y, y') \wedge \forall d, e\ (y' S_x dRe \to yRd)).$$

An immediate consequence of Lemma 6.11 is the following theorem.

Theorem 6.14. *For F a finite frame, we have*

$$F \models \mathsf{B} \quad \Leftrightarrow \quad F \models \mathcal{C}_{\mathsf{B}}.$$

Note that the M-frame condition can be seen as a special case of the frame condition of B: we demand that $\mathcal{S}$ be the identity relation.

It is not hard to see that the frame condition of M_0 follows from $\mathcal{C}_0$. And indeed, $\mathbf{IL}\mathsf{B} \vdash \mathsf{M}_0$ as $\Diamond A \in \mathsf{ES}_2$ and $A \rhd B \to \Diamond A \rhd B$. Actually, we have that $\mathbf{IL}\mathsf{B}_1 \vdash \mathsf{M}_0$.

7 Beklemishev and Zambella

Zambella proved in [18] a fact concerning Π_1-consequences of theories with a Π_2 axiomatization. As we shall see, his result has some repercussions on the study of the interpretability logic of PRA.

Lemma 7.1 (Zambella)**.** *Let T and S be two theories axiomatized by Π_2-axioms. If T and S have the same Π_1-consequences then $T + S$ has no more Π_1-consequences than T or S.*

In [18], Zambella gave a model-theoretic proof of this lemma. As was sketched by G. 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 Π_1-conservativity logic of Π_2-axiomatized theories. We denote it here as Z^{Π_1}.

$$\mathsf{Z}^{\Pi_1} \quad (A \equiv_{\Pi_1} B) \to A \rhd_{\Pi_1} A \wedge B \quad \text{for } A \text{ and } B \text{ in } \mathsf{EP}_2^{\mathsf{c}}.$$

Since PRA is Π_2 axiomatized and proves totality of the supexp function this principle applies to PRA. But there are repercussions for the interpretability logic of PRA as well. We know that for reflexive theories Π_1-conservativity coincides with interpretability. We also know that any Σ_2-extension of PRA is reflexive (Lemma 2.1). Altogether this means that a statement $\alpha \rhd \beta$ and $\alpha \rhd_{\Pi_1} \beta$ are equivalent if α is in Σ_2 and $\mathrm{PRA} + \alpha$ is Π_2-axiomatized, i.e. α is in Δ_2.

We arrive at Zambella's principle for interpretability logic:

$$\mathsf{Z} \quad (A \equiv B) \to A \rhd A \wedge B \quad \text{for } A \text{ and } B \text{ in } \mathsf{ED}_2$$

For the Π_1-conservativity logic of PRA, the principle Z^{Π_1} is really informative (see [3]), it is the only principle know on top of the basic ones for the Π_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 $\mathbf{IL}\mathsf{B}$ 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 $\mathbf{IL}\mathsf{B}_0 \vdash \mathsf{Z}$, B_0 being a restriction of B to ED_2 formulas, see Definition 6.8. The proof in [11] of the same fact was not correct.

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

Definition 7.2. 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 $\pm C_1 \wedge \ldots \wedge \pm C_n$ 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 ED_2-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 subforumulas maximal w.r.t. box-depth).

Theorem 7.3. $\mathbf{IL}\mathsf{B}_0 \vdash \mathsf{Z}$

Proof. Let $A, B \in \mathsf{ED}_2$ 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 $A \rhd A \wedge B$. Since A comes in full DNF, this means to show, for each disjunct D of A, that $D \rhd A \wedge B$. 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 $D^{\Box}$ of boxed formulas occurring positively in it. We shall write $D^{\Box}$ also for the conjunction of its members.

We first show, if D is a member of A or B which has a maximal set $D^{\Box}$ (no disjunct E with $E^{\Box}$ properly containing $D^{\Box}$ occurs in A or B) then $D \rhd A \wedge B$:

Suppose such D is in A, the other case is symmetrical. Since $D \rhd A$ we have also $D \rhd B$. Then, noting that $D^{\Box}$ is a conjunction of boxed formulas and applying B_0, we obtain $D \rhd B \wedge D^{\Box}$.

Now take any disjunct E of B for which $E^{\Box}$ does not contain $D^{\Box}$. Then E contradicts $D^{\Box}$ by its negative part. We distinguish two cases: if for all E in B the set $E^{\Box}$ does not contain $D^{\Box}$, then B contradicts $D^{\Box}$. It follows from $D \rhd B \wedge D^{\Box}$ that $D \rhd \bot$. Then clearly $D \rhd A \wedge B$.

Otherwise B does contain E with $E^{\Box}$ containing $D^{\Box}$. But since D has a maximal Box-set, E and D must be the same and D occurs in B as well. Thus $D \rhd B \wedge D$ and, since $\vdash D \rightarrow A$, also $D \rhd A \wedge B$.
We have shown that all maximal disjuncts interpret $A \wedge B$.

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 $m \geq k' > k$ and all disjuncts in either A or B with $D^{\Box}$ of size k', $D \rhd A \wedge B$ (this has been already shown for k equal the size of the maximal Box-set in A and B certainly less then m). Consider a disjunct D of A, the other case is again symmetrical. Assume w.l.o.g. that $D^{\Box}$ has size k. We have to show $D \rhd A \wedge B$:

Since $D \rhd A$ and hence $D \rhd B$, we again have that $D \rhd B \wedge D^{\Box}$. Now $D^{\Box}$ conflicts with all the disjuncts of B Box-set of which is not a superset of $D^{\Box}$. Again, we distinguish two cases: if there are none disjuncts of B with a Box-set which is a superset of $D^{\Box}$ then B conflicts with $D^{\Box}$ and $D \rhd \bot$ and thus $D \rhd A \wedge B$.

Otherwise some disjuncts of B do have a Box-set which is a superset of $D^{\Box}$. Let $E_1, \ldots, E_l$ be all such disjuncts of B. Then, since $D \rhd B \wedge D^{\Box}$ and $\vdash B \wedge D^{\Box} \rightarrow E_1 \vee \ldots \vee E_l$ ($E_1 \vee \ldots \vee E_l$ is the part of B not conflicting with $D^{\Box}$), we obtain $D \rhd E_1 \vee \ldots \vee E_l$. Now it suffices to show that each E_i interprets $A \wedge B$. But this is the induction hypothesis since all $E_i^{\Box}$ have size greater then k. $\Box$

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_2 are just the propositional combinations of $\Box$-formulas. Now let us allow in A, B not only $\Box$-formulas but also formulas of the form $C \rhd D$ with $C, D \in \mathsf{ES}_2$. Let us furthermore write $\boxdot F$ for $F \wedge \Box F$.

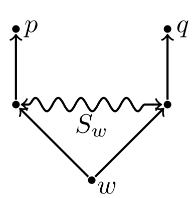
Then, if $C_1 \rhd D_1, \ldots, C_n \rhd D_n$ are all the formulas in either A or B of the above mentioned form and $E_1, \ldots E_n$ are arbitrary formulas in ES_2, then

$$\boxdot(((C_1 \rhd D_1) \leftrightarrow E_1) \wedge \cdots \wedge ((C_n \rhd D_n) \leftrightarrow E_n)) \rightarrow ((A \equiv B) \rightarrow A \rhd A \wedge B)$$

can still be seen to be valid using the same considerations that led us to the principle Z in the first place.

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) $\subseteq$ **IL**$\mathsf{M} \cap$ **IL**P and Z is not in **IL**P: consider the following model:



We have $w \Vdash \Diamond p \equiv \Diamond q$ and $w \nVdash p \rhd p \wedge q$, thus Zambella fails. The model is clearly an **IL**P model.

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 **IL**M since M is not in **IL**(PRA), as A. Visser discusses in [17]: the two logics cannot be the same because if **IL**M 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 **IL**M which is not in **IL**(PRA) can be given:

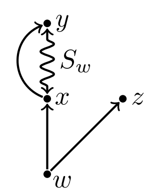
$$A \rhd \Diamond B \to \Box(A \rhd \Diamond B).$$

That this formula is not in **IL**(PRA) can be shown using Shavrukov's result from [14] about complexity of the set $\{\psi | \psi \in \Pi_1 \ \& \ \phi \rhd \psi\}$; see [17] for the full proof.

We know that M_0 is provable in **IL**B. 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.

W **vs.** B**:** It is easy to see that $\mathsf{B} \nvdash \mathsf{W}$ since the later is in **IL**(All) while the former is not in it.

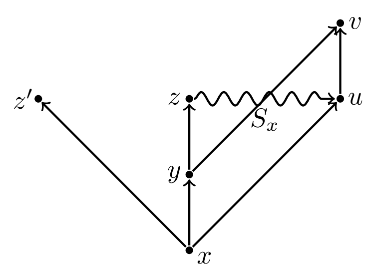
The following frame



is an **IL**B frame and it violates the frame condition for W: $wRxRy$ and xS_wyS_wx and wRz. Now z is bi-similar to y and B is ensured.

R **vs.** B**:** Again, since $\mathsf{R} \in$ **IL**(All), it cannot be that $\mathsf{R} \vdash \mathsf{B}$.

The following frame



is an **IL**B-frame violating the frame condition of R:

We have a basic situation violating R, $xRyRzS_xuRv$ and $\neg zS_yv$. To ensure B for y we add an arrow yRv, to ensure B for z, we add a bi-similar world z' such that xRz' and z' has no successors at all.

R **vs.** W**:** already discussed in [8].

It is clear from our exposition that, though we have solved a number of problems concerning **IL**(PRA), many remain open, e.g. those connected with our incomplete knowledge of **IL**(All). Also, we lack a modal completeness theorem for **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 interpetability is far from being a finished subject.

References

[1] L.D. Beklemishev, *Induction Rules, Reflection Principles, and Provably Recursive Functions*, Annals of Pure and Applied Logic **85** (1997), 193–242.

[2] ———, *Proof-Theoretic Analysis by Iterated Reflection*, Archive for Mathematical Logic **42** (2003), no. 6, 515–552.

[3] L.D. Beklemishev and A. Visser, *On the Limit Existence Principles in Elementary Arithmetic and related topics*, Logic Group Preprint Series 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 and G. Japaridze, *The Logic of Provability*, Handbook of Proof Theory. Studies in Logic and the Foundations of Mathematics, Vol.137. (S.R. Buss, ed.), Elsevier, Amsterdam, 1998, pp. 475–546.

[6] D.H.J. de Jongh and F. Veltman, *Provability Logics for Relative Interpretability*, [12], 1990, pp. 31–42.

[7] ———, *Modal Completeness of IL*W, Essays dedicated to Johan van Benthem on the occasion of his 50th birthday (J. Gerbrandy, M. Marx, M. Rijke, and Y. Venema, eds.), Amsterdam University Press, Amsterdam, 1999.

[8] E. Goris and J.J. Joosten, *Modal Matters in Interpretability Logics*, Logic Group Preprint Series 226, University of Utrecht, March 2004.

[9] ———, *A new principle in the interpretability logic of all reasonable arithmetical theories*, IGPL (2008).

[10] P. Hájek and F. Montagna, *The logic of* Π_1*-Conservativity*, Archiv für Mathematische Logik und Grundlagenforschung **30** (1990), 113–123.

[11] J.J. Joosten, *Interpretability Formalized*, Ph.D. thesis, Utrecht University, November 2004, ISBN: 90-393-3869-8.

[12] P.P. Petkov (ed.), *Mathematical Logic, 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 (in Russian)*, Tech. Report Report No.5, Steklov Mathematical Institute, Moscow, 1988.

[14] V.Yu. Shavrukov, *Interpreting Reflexive Theories in Finitely many Axioms*, Fundamenta Mathematicae **152** (1997), 99–116.

[15] A. Tarski, A. Mostowski, and R. Robinson, *Undecidable Theories*, North–Holland, Amsterdam, 1953.

[16] A. Visser, *Interpretability Logic*, [12], 1990, pp. 175–209.

[17] A. Visser, *An Overview of Interpretability Logic*, Advances in modal logic '96 (M. Kracht, M. de Rijke, and H. Wansing, eds.), 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.