The many faces of interpretability

J.J.Joosten[1] and E.Goris[2]

[1] University of Amsterdam
[2] City University of New York

Abstract. In this paper we discus work in progress on interpretability logics. We show how semantical considerations have allowed us to formulate non-trivial principles about formalized interpretability. In particular we falsify the conjecture about the nature of the interpretability logic of all reasonable arithmetical theories. We consider this an interesting example of how purely semantical considerations give new non-trivial facts about syntactical and arithmetical notions.
In addition we give some apparatus that allows us to push 'global' semantical properties into more 'local' syntactical ones. With this apparatus, the rather wild behavior of the different interpretability logics are nicely formulated in a single notion that expresses their differences in a uniform way.

This paper consists of three parts. We start of by giving a short introduction to interpretability logics. In the second part we discuss how a careful analysis of the modal semantical behaviour of interpretability logics lets us formulate non-trivial interpretability principles. In the third part we present a semantical bookkeeping tool which pushes 'global' semantical considerations into 'local' syntactical ones. The hope is that this machinery will provide a general and uniform treatment of the bewildering field of modal interpretability logics.

1 Introduction

All techniques and results in this paper revolve around the notions of interpretations between formal theories, and of formalized interpretability. We shall not give a detailed definition here of interpretations and confer the reader to, e.g., [1].

Roughly, an interpretation j of one theory T into another theory S, we write $j : S \rhd T$, is a structure preserving map, mapping theorems φ of T to theorems φ^j of S. Structure preserving means as much as commuting with proof constructions and with logical connectives. Thus, for example, the map that sends any formula in the language of T to a some tautology in the language of S does in general not qualify as an interpretation.

If there exists an interpretation $j : S \rhd T$, we say that S interprets T, or that T is interpretable in S and write $S \rhd T$. In a sense, if $S \rhd T$ we can say that the theory S is at least as strong as the theory T. In a sense, interpretability can be seen as a generalization of provability. Certainly, if S proves all theorems of T, then via the identity mapping we get that $S \rhd T$.

Interpretations turn up time and again in any part of mathematics or metamathematics. They have been used to establish undecidability of formal theories [2], relative consistency results [3] [4], incompleteness results [5] and also are used in foundational studies [6], [7], [8]. In this paper we shall only be dealing with interpretations between first order theories in which some minimal part of number theory "lives" in a natural way. One way to study the behaviour of interpretability is by employing so-called interpretability logics.

1.1 Interpretability logics

With the 'logics approach' we can capture a large part of the structural behavior of interpretations. Let us consider such a structural rule.

For any theories U, V and W we have that, if $U \rhd V$ and $V \rhd W$, then also $U \rhd W$. It is not hard to catch this in a modal logic. But modal logics talk about propositions and interpretability talks about theories.

It does not seem to be a good idea to directly translate propositional variables to theories. For what does the negation of a theory mean? And how to read implication? And how to translate modal statements involving iterated modalities?

The usual way to relate modal logics to interpretability is to translate propositional variables to arithmetical sentences that are added to some base theory T. Of course, the meta-theory should be strong enough to allow for arithmetization. By this approach we get quite an expressive formalism in which the logic of provability is naturally embedded.

We shall work with a modal language containing two modalities, a unary modality $\Box$ and a binary modality $\rhd$. As always, we shall use $\Diamond A$ as short for $\neg\Box\neg A$. Apart from propositional variables we also have two constants $\top$ and $\bot$ in our language.

In this paper we thus use the same symbol $\rhd$ both for formalized interpretability and for our binary modal operator. The same holds for $\Box$. But the context will always decide on how to read the symbol.

In writing formulas we shall omit brackets that are superfluous according to the following reading conventions. We say that the operators $\Diamond$, $\Box$ and $\neg$ bind equally strong. They bind stronger than the equally strong binding $\wedge$ and $\vee$ which in turn bind stronger than $\rhd$. The weakest (weaker than $\rhd$) binding connectives are $\rightarrow$ and $\leftrightarrow$. We shall also omit outer brackets. Thus, we shall write $A \rhd B \rightarrow A \wedge \Box C \rhd B \wedge \Box C$ instead of $((A \rhd B) \rightarrow ((A \wedge (\Box C)) \rhd (B \wedge (\Box C))))$.

Definition 1 (Arithmetical realization). *An arithmetical T-realization is a map $*$ sending propositional variables p to arithmetical sentences p^*. The realization $*$ is extended to a map that is defined on all modal formulae as follows.*

It is defined to commute with all boolean connectives. Moreover $(A \rhd B)^* = (T \cup \{A^*\}) \rhd (T \cup \{B^*\})$ (we shall write $A^* \rhd_T B^*$) and $(\Box A)^* = \Box_T A^*$. Here $\rhd_T$

and $\Box_T$ *denote the formulas expressing formalized interpretability and formalized provability respectively, over* T*.*

We shall reserve the symbol $*$ to range over T-realizations. Moreover, we will speak just of realizations if the T is clear from the context. In the literature realizations are also referred to as interpretations or translations. As these words are already reserved for other notions in our paper, we prefer to talk of realizations.

Definition 2 (Interpretability principle, Interpretability logic). *A modal formula* A *is an* interpretability principle *of a theory* T*, if* $\forall * T \vdash A^*$*. The* interpretability logic *of a theory* T*, we write* $\mathbf{IL}(\mathrm{T})$*, is the set of all the interpretability principles of* T *or a logic that generates it.*

Definition 3 (IL). *With* IL *we will refer to the logic axiomatized by classical propositional logic, the following set of axiom schemata*

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

using necessitation (from A*, conclude* $\Box A$*) and modus ponens.*

With ILX we denote the logic that arises by adding a principle X to the axiom schemata of IL and likewise for adding more principles. In this paper we shall discuss the following principles.

$$\begin{array}{ll}
\mathsf{W} & := A \rhd B \to A \rhd B \wedge \Box\neg A \\
\mathsf{M} & := A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C \\
\mathsf{P} & := A \rhd B \to \Box(A \rhd B) \\
\mathsf{M_0} & := A \rhd B \to \Diamond A \wedge \Box C \rhd B \wedge \Box C \\
\mathsf{R} & := A \rhd B \to \neg(A \rhd \neg C) \rhd B \wedge \Box C
\end{array}$$

1.2 The quest for IL(All)

By a result of Shavrukov [9] and independently, Berarducci [10], it is known that $\mathbf{IL}(\mathrm{T})$=ILM whenever T is an essentially reflexive theory (a theory that proves the consistency of all its finite subtheories).

By a result of Visser [11] it is known that $\mathbf{IL}(\mathrm{T})$=ILP whenever T is a finitely axiomatizable theory of some minimal strength.

We see here a phenomenon different from provability logics: different theories can have different interpretability logics. We are interested in the collection of modal formulae that are interpretability principles of any reasonable theory. Of course, the term *reasonable* should be agreed upon. It is good to start out with a very general notion, that is, putting very little constraints. As in [12] we just demand that we can do basic syntactic operation.[3]

Definition 4. *The interpretability logic of all reasonable arithmetical theories, we write* $\mathbf{IL}(\mathrm{All})$*, is the set of formulas* φ *such that* $\forall T \, \forall * \;\; T \vdash \varphi^*$*. Here the* T *ranges over all the reasonable arithmetical theories.*

Throughout the previous decades many a conjecture has been made as to the nature of $\mathbf{IL}(\mathrm{All})$, but up to the date of today the problem remains unsettled.

1.3 Logics, semantics and interactions

Interpretability logics come with a natural Kripke semantics. Below we define this semantics. The logic IL is known to be complete with respect to this semantics [1] and similar completeness results have been obtained for ILP, ILM, $\mathsf{ILM_0}$ and ILW [11][10][9][14][15].

Definition 5 (Veltman Frame). *A* Veltman frame*, or just* frame*, is a triple* $F = \langle W, R, S\rangle$ *where*

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

Definition 6 (Veltman model). *A* Veltman model*, or simply* model*, is a quadruple* $M = \langle W, R, S, \Vdash\rangle$ *where* $\langle W, R, S\rangle$ *is a Veltman frame and* $\Vdash$ *is a (forcing) relation between elements of* W *and* IL*-formulas satisfying the following requirements.*

1. $w \Vdash A \rhd B$ *iff for each* u *such that* wRu *and* $u \Vdash A$*, there exists* uS_wv *such that* $v \Vdash B$
2. $w \Vdash \Box A$ *iff for each* u *such that* wRu *we have* $u \Vdash A$
3. $\Vdash$ *commutes with boolean connectives, e.g.* $w \Vdash A \wedge B$ *iff* $w \Vdash A$ *and* $w \Vdash B$

As always, the modal semantics provide a good heuristics for finding modal proofs. On the other hand, there also is a deep connection between modal models, and models of arithmetic. For example, one can consider the class of models of PA and define accessibility relations between these models in such a way that, in a sense, Veltman semantics arises. Thus, nodes of a Veltman model correspond

[3] This can be expressed by $T \rhd \mathsf{S}^1_2$ where S^1_2 is Buss's theory of bounded arithmetic. See, e.g., [13].

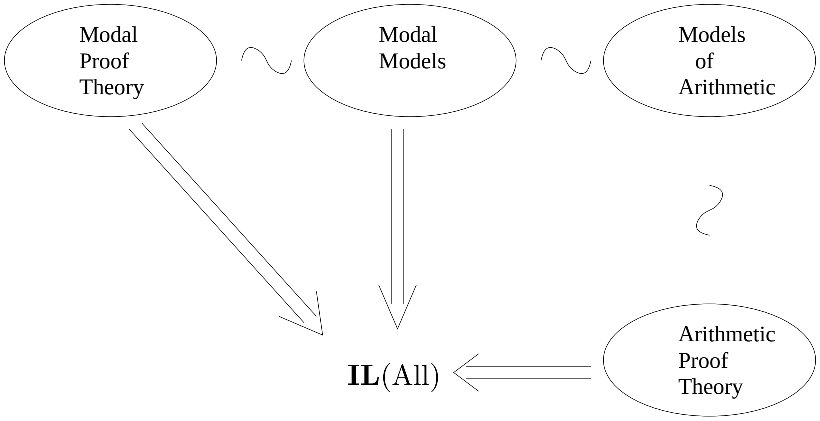

Fig. 1. Many disciplines interacting

to models of PA and the forcing relation $\Vdash$ is just first order model theoretical satisfaction $\models$. (See Appendix D from [16].)

Although there are some indications[4] as how to generalize these results to models of reasonable arithmetic, this project remains open. Nevertheless, the intuitions on the arithmetical models and the correspondence to Veltman models, are still usable and yield good heuristics for work on modal models. We shall see in the next section how the modal models contribute in a very fruitful way to finding new principles in **IL**(All).

Of course, also direct arithmetic consideration of proof theoretic nature can yield new insights and possibly new principles for **IL**(All). The latest progress however comes primarily from semantical, i.e., modal considerations. In Figure 1 we schematically represent the various fields and interactions that come together when studying the logic **IL**(All).

2 Veltman semantics and the road to IL(All)

As already mentioned above, most principles about interpretability for fragments of arithmetic are direct consequences of proof theoretic facts about these fragments. For example the Orey-Hájek characterization of interpretability for essentially reflexive theories directly translates to the M scheme:

$$\mathsf{M}: \quad A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C.$$

Roughly, the Orey-Hájek characterization for reflexive theories gives us that if $A \rhd B$, then we can make for any model M of A an internally definable model

[4] A model theoretic characterization of interpretability along the lines of [12], Section 2.3 is certainly relevant here.

M' of B which is an end extension of M. Clearly, $\Box C$, being a Σ_1-formula, is preserved under the end extension. This is a nice illustration of how reasoning in terms of models provides a good heuristics.

If we consider the notion of interpretability for finitely axiomatizable fragments of arithmetic, it is not hard to see that interpretability is actually formalizable by a Σ_1-formula. And thus we immediately get, by provable Σ_1-completeness, the P scheme:

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

Moving along such proof theoretic lines, one can justify W and $\mathsf{M_0}$ as well.

There is another way to come up with valid principles that isn't fully justified in itself (it is more of a heuristic rather than a method). But this has nevertheless given some good results and is much lighter than the heavy proof theoretic machinery mentioned above.

2.1 Arithmetical Principles from Modal Semantics

Like with other modal logics, for interpretability logics there exists a well understood frame correspondence theory. It is not hard to see that the scheme M holds on any frame F (that is, holds on any model with underlying frame F) if and only if $uS_xvRw \to uRw$. Likewise we can formulate a frame condition for P being $xRyRuS_xv \to uS_yv$.

A principle in **IL**(All) must certainly be an interpretability principle for essentially reflexive theories and also for finitely axiomatizable theories. Thus, the principle should hold on any ILM-frame and also on any ILP-frame. Hence, modal semantics severely confines the search space for new principles. We can just look for 'natural' semantic conditions that hold both on ILM and on ILP frames, and then look for the corresponding principle. Heuristics for the 'naturality' are provided by the earlier described relations with arithmetical models, arithmetical proof theory, modal logics and syntactical form.

2.2 From $\mathsf{M_0}$ to $\mathsf{P_0}$, and from $\mathsf{P_0}$ to R

The scheme $\mathsf{M_0}$ corresponds to the frame condition

$$wRxRyS_wy'Rz \text{ implies } xRz.$$

Stagnation in attempts to prove modal completeness of IL$\mathsf{M_0}$ led in [17] to proposing a stronger scheme

$$wRxRyS_wy'Rz \text{ implies } xRz \wedge yS_xz,$$

and the corresponding principle $\mathsf{P_0}$

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

which is easily seen to be valid under this stronger frame condition. Using methods from [18] one easily shows that this principle is arithmetically valid. However, careful analysis of what is needed to show modal completeness of ILP_0 revealed (see[5] [14]) the scheme R

$$A \rhd B \to \neg(A \rhd C) \rhd B \wedge \Box\neg C.$$

In what follows we show that R does not follow from P_0. In fact we show that $\mathrm{ILP}_0\mathsf{M}_0\mathsf{W}$ is incomplete. Let us first calculate the frame condition of R. It turns out to be the same frame condition as for P_0 (see [17]).

Lemma 7. $F \models \mathsf{R} \Leftrightarrow [xRyRzS_xuRv \to zS_yv]$

Proof. "$\Leftarrow$" Suppose that at some world $x \Vdash A \rhd B$. We are to show $x \Vdash \neg(A \rhd \neg C) \rhd B \wedge \Box C$. Thus, if $xRy \Vdash \neg(A \rhd \neg C)$ we need to go via an S_x to a u with $u \vdash B \wedge \Box C$.

As $y \Vdash \neg(A \rhd \neg C)$, we can find z with $yRz \Vdash A$. Now, by $x \Vdash A \rhd B$, we can find u with $yS_xu \Vdash B$. We shall now see that $u \Vdash B \wedge \Box C$. For, if uRv, then by our assumption, zS_yv, and by $y \Vdash \neg(A \rhd \neg C)$, we must have $v \vdash C$. Thus, $u \Vdash B \wedge \Box C$ and clearly yS_xu.

"$\Rightarrow$" We suppose that R holds. Now we consider arbitrary a, b, c, d and e with $aRbRcS_adRe$. For propositional variables p, q and r we define a valuation $\Vdash$ as follows.

$$\begin{aligned}x \Vdash p &:\Leftrightarrow x = c\\ x \Vdash q &:\Leftrightarrow x = d\\ x \Vdash r &:\Leftrightarrow cS_bx\end{aligned}$$

Clearly, $a \Vdash p \rhd q$ and $b \Vdash \neg(p \rhd \neg r)$. By R we conclude $a \Vdash \neg(p \rhd \neg r) \rhd q \wedge \Box r$. Thus, $d \Vdash q \wedge \Box r$ which implies cS_be.

Theorem 8. $\mathrm{ILP}_0\mathsf{M}_0\mathsf{W} \nvdash \mathsf{R}$

Proof. We consider the model M from Figure 2 and shall see that $M \models \mathrm{ILP}_0\mathsf{M}_0\mathsf{W}$ but $M, a \nVdash \mathsf{R}$. Since forcing of formulas in a model is preserved under modal derivability, we conclude that $\mathrm{ILP}_0\mathsf{M}_0\mathsf{W} \nvdash \mathsf{R}$.

As M satisfies the frame condition for $\mathsf{M}_0\mathsf{W}$, it is clear that $M \models \mathsf{M}_0\mathsf{W}$. We shall now see that $M \models A \rhd \Diamond B \to \Box(A \rhd B)$ for any formulas A and B.

A formula $\Box(A \rhd B)$ can only be false at some world with at least two successors. Thus, in M, we only need to consider the point a. So, suppose $A \rhd \Diamond B$. For which x with aRx can we have $x \Vdash A$?

As we have to be able to go via an S_x-transition to a world where $\Diamond B$ holds, the only candidates for x are b, c and d. But clearly, c and f make true the same modal formulas. From f it is impossible to go to a world where $\Diamond B$ holds.

Thus, if $a \Vdash A \rhd \Diamond B$, the A can only hold at b or at d. But this automatically implies that $a \Vdash \Box(A \rhd B)$ and $M \models \mathsf{P}_0$.

It is not hard to see that $a \nVdash \mathsf{R}$. Clearly, $a \Vdash p \rhd q$ and $b \Vdash \neg(p \rhd \neg r)$. However, $d \nVdash q \wedge \Box r$ and thus $a \nVdash \neg(p \rhd \neg r) \rhd q \wedge \Box r$.

[5] In this reference a slightly different principle was formulated. In [19] the principle R was first published.

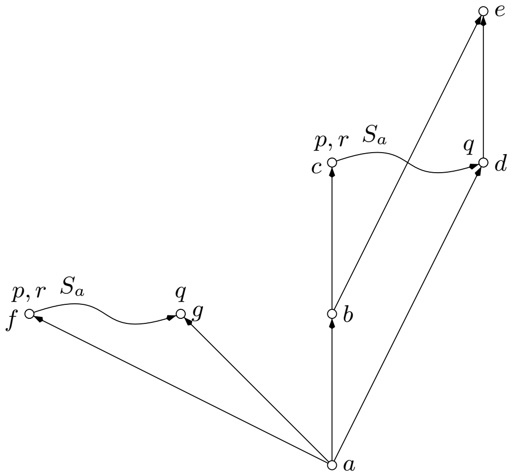

Fig. 2. ILP$_0$M$_0$W is incomplete

The following lemma tells us that ILR is a proper extension of ILM$_0$P$_0$.

Lemma 9. ILR $\vdash$ M$_0$, P$_0$

Proof. As IL $\vdash \Diamond A \wedge \Box C \to \neg(A \rhd \neg C)$ we get that $A \rhd B \to \Diamond A \wedge \Box C \rhd \neg(A \rhd \neg C)$ and M$_0$ follows from R. The principle P$_0$ follows directly from R by taking $C = \neg B$.

2.3 From R to, who knows

By systematically searching the "natural" frame conditions that live in the intersection of ILP and ILM one can find new candidate principles. The authors claim to recently have found a new whole "hierarchy" of principles in **IL**(All).

3 Full Labels

In this section we address difficulties found in modal completeness proofs not by formulating new principles, but by a book keeping apparatus which allows us to 'localize' certain 'global' desirable properties of models. It has been put to work successfully in [19] where a considerable simplification of the modal completeness proof of ILW (as given in [15]) is given. Here we shall give some basic theory.

As is in general the case with model constructions, maximal consistent sets of formulas play an important role. In addition, in model constructions for interpretability logics, a central notion in these constructions is the notion of a critical successor [20].

Definition 10 (Critical successor). *Let Γ and Δ be maximal consistent sets of formulas and let B be a formula. We say that Δ is a B-critical successor of*

Γ, and write $\Gamma \prec_B \Delta$ when if $A \rhd B \in \Gamma$, then $\neg A, \Box\neg A \in \Delta$. Moreover, for some $\Box C$ we have $\Box C \in \Delta$ but $\Box C \notin \Gamma$.

In this section we will expose a generalization of critical successor and show how it can be used to solve, in a uniform way certain problematic aspects of modal completeness proofs.

Definition 11 (Assuring successor). *Let S be a set of formulas. We define $\Gamma{\prec_S}\Delta$, and say that Δ is an S-*assuring successor *of Γ, if for any finite $S' \subseteq S$ we have $A \rhd \bigvee_{S_j \in S'} \neg S_j \in \Gamma \Rightarrow \neg A, \Box\neg A \in \Delta$ and for some $\Box C \in \Delta$ we have $\Box C \notin \Gamma$.*

Lemma 12. *For the relation $\prec_S$ we have the following observations.*

1. *$\Gamma{\prec_\emptyset}\Delta \Leftrightarrow \Gamma \prec \Delta$*
2. *Δ is a B-critical successor of Γ $\Leftrightarrow$ $\Gamma{\prec_{\{\neg B\}}}\Delta$*
3. *$S \subseteq T$ & $\Gamma{\prec_T}\Delta \Rightarrow \Gamma{\prec_S}\Delta$*
4. *$\Gamma{\prec_S}\Delta \prec \Delta' \Rightarrow \Gamma{\prec_S}\Delta'$*
5. *$\Gamma{\prec_S}\Delta \Rightarrow S, \Box S \subseteq \Delta, \Diamond S \subseteq \Gamma$ and for all A, $\Diamond A \notin S$*

Theorem 13. *Let Γ be a* MCS *and S a set of formulas. If for any choice of $S_i \in S$ we have that $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$, then*[6] *there exists a* MCS *Δ such that $\Gamma{\prec_S}\Delta \ni B, \Box\neg B$.*

Proof. Suppose for a contradiction there is no such Δ. Then there is a formula[7] A such that for some $S_i \in S$, $(A \rhd \bigvee \neg S_i) \in \Gamma$ and $\Box\neg B, B, \Box\neg A, \neg A \vdash \bot$. Then $\vdash \Box\neg B \wedge B \rhd A \vee \Diamond A$ and we get $\vdash B \rhd A$. As $(A \rhd \bigvee \neg S_i) \in \Gamma$, also $(B \rhd \bigvee \neg S_i) \in \Gamma$. A contradiction.

Lemma 14. *Let Γ be a* MCS *such that $\neg(B \rhd C) \in \Gamma$. Then there is a* MCS *$\Delta$ such that $\Gamma{\prec_{\{\neg C\}}}\Delta$ and $B, \Box\neg B \in \Delta$.*

Proof. Taking $S = \{\neg C\}$ in Theorem 13.

Lemma 15. *Let Γ and Δ be* MCS*'s such that $A \rhd B \in \Gamma{\prec_S}\Delta \ni A$. Then there is a* MCS *$\Delta'$ such that $\Gamma{\prec_S}\Delta' \ni B, \Box\neg B$.*

Proof. First we see that for any choice of S_i, $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$. Suppose not. Then for some S_i, $(B \rhd \bigvee \neg S_i) \in \Gamma$ because Γ is a MCS. But then $(A \rhd \bigvee \neg S_i) \in \Gamma$ and by $\Gamma{\prec_S}\Delta$ we have $\neg A \in \Delta$. A contradiction. So $\neg(B \rhd \bigvee \neg S_i) \in \Gamma$ for any choice of S_i and we can apply Theorem 13.

Lemmata 14, 15 are the obvious generalizations of the corresponding lemmata involving criticality instead of assuringness. To clarify the benefits of assuringness over criticality let us roughly identify the three main points when building a counter model $\langle W, R, S, V\rangle$ for some unprovable formula (in some extension of IL). We take W a multi set of MCS's and build the model in a step by step fashion.

[6] It is easy to see that we actually have iff.

[7] By compactness there are finitely many A_j with for some S_i^j, $(A_j \rhd \bigvee \neg S_i^j) \in \Gamma$ and $\Box\neg B, B\neg A_j, \Box\neg A_j \vdash \bot$. We can take A to be $\bigvee_j A_j$.

1. For each $\Gamma \in W$ with $\neg(A \rhd B) \in \Gamma$ we should add some B-critical successor (equivalently $\{\neg B\}$-assuring successor) Δ to W for which $A \in \Delta$.
2. For each $\Gamma, \Delta \in W$ with $C \rhd D \in \Gamma R \Delta \ni C$ we should add a Δ' to W for which $\Gamma \prec \Delta' \ni D$. Moreover if Δ is a B-critical successor of Γ then then we should be able to choose Δ' a B-critical successor of Γ as well.
3. We should take care of the frame conditions.

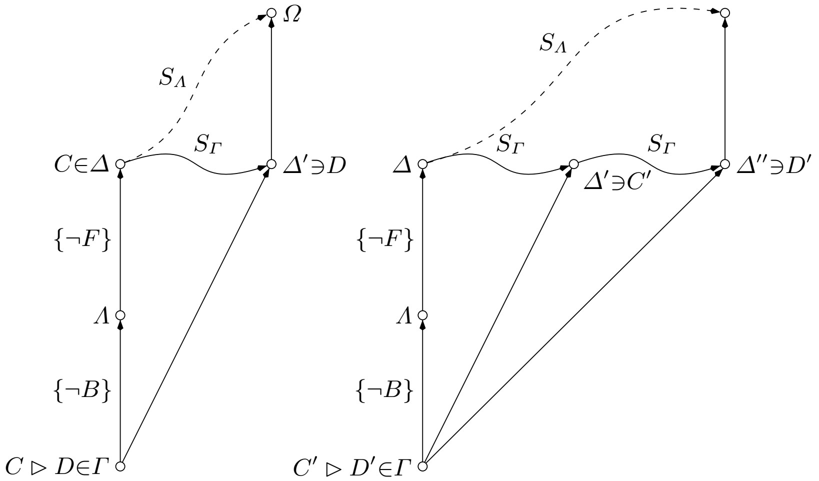

Fig. 3. R frame condition

When working in IL, Lemma 14 handles Item 1 and Lemma 15 handles Item 2. Making sure that the frame conditions are satisfied does not impose any problems [1]. With extensions of IL the situation regarding the frame conditions becomes more complicated [20][14]. Let us clarify this by looking at ILR. The additional frame condition is as follows [14][8].

$$wRxRyS_wy'Rz \Rightarrow yS_xz$$

This is depicted in the leftmost picture in Figure 3. Let us use the notation as in Item 2: Δ' was added to the model since $C \rhd D \in \Gamma R \Delta \ni C$. Since Δ lies F-critical (equivalently $\{\neg F\}$-assuring) above Λ, we should not only make sure that Δ' lies B critical above Γ, but also that for any successor Ω of Δ' lies F critical above Λ.

One way to guarantee this is to actually require that $\Box\neg H \in \Delta'$ whenever $H \rhd F \in \Lambda$. As one easily checks, it is quite easy to prove such a Lemma in

[8] In [14] the modal principle $A \rhd B \to \neg(A \rhd \neg C) \wedge (D \rhd C) \rhd B \wedge \Box C$ was called R. This principle and the one called R in this paper are easily seen to be equivalent over IL.

ILR but we have oversimplified[9] the situation. Consider the rightmost picture in Figure 3. That is, after having added $\varDelta'$ to the model we are required to add some $\varDelta''$ with $D' \in \varDelta'$ to the model since $C' \rhd D' \in \varGamma$ and $C' \in \varDelta'$. By the transitivity of $S_{\varGamma}$ we require that $\Box\neg H \in \varDelta''$ whenever $H \rhd F \in \varLambda$. In this situation it is not so clear what to do.

Although for ILM$_0$ [14] and ILW [15] there where add hoc solutions to similar problems, criticality seemed too weak a notion for a more uniform solution. As the lemmata below will show, assuringness does give us a uniform method for handling these kind of situations.

In what follows put, for any set of formulas T,

$$\varDelta^{\Box}_{T} = \{\Box\neg A \mid T' \subseteq T \text{ finite }, A \rhd \bigvee_{T_i \in T'} \neg T_i \in \varDelta\},$$
$$\varDelta^{\boxdot}_{T} = \{\Box\neg A, \neg A \mid T' \subseteq T \text{ finite }, A \rhd \bigvee_{T_i \in T'} \neg T_i \in \varDelta\}.$$

Lemma 16. *For any logic (i.e. extension of* IL*) we have* $\varGamma \prec_S \varDelta \Rightarrow \varGamma \prec_{S \cup \varGamma^{\boxdot}_{S}} \varDelta$.

Proof. Suppose $\varGamma \prec_S \varDelta$ and $C \rhd \bigvee \neg S_i \vee \bigvee A_j \vee \Diamond A_j \in \varGamma$. Then $C \rhd \bigvee \neg S_i \vee \bigvee A_j \in \varGamma$ and thus $C \rhd \bigvee \neg S_i \vee \bigvee \neg S^j_k \in \varGamma$ which implies $\neg C, \Box\neg C \in \varDelta$.

Lemma 17. *For logics containing* M *we have* $\varGamma \prec_S \varDelta \Rightarrow \varGamma \prec_{S \cup \varDelta^{\Box}_{\emptyset}} \varDelta$.

Proof. Note that $\varDelta^{\Box}_{\emptyset} = \{\Box C \mid \Box C \in \varDelta\}$. We consider A such that for some $S_i \in S$ and $\Box C_j \in \varDelta^{\Box}_{\emptyset}$, $(A \rhd \bigvee \neg S_i \vee \bigvee \neg\Box C_j) \in \varGamma$. By M, $(A \wedge \bigwedge \Box C_j \rhd \bigvee \neg S_i) \in \varGamma$, whence $\boxdot\neg(A \wedge \bigwedge \Box C_j) \in \varDelta$. As $\bigwedge \Box C_j \in \varDelta$, we conclude $\neg A, \Box\neg A \in \varDelta$.

Lemma 18. *For logics containing* P *we have* $\varGamma \prec_S \varLambda \prec_T \varDelta \Rightarrow \varGamma \prec_{S \cup \varLambda^{\boxdot}_{T}} \varDelta$.

Proof. Suppose $C \rhd \bigvee \neg S_i \vee \bigvee A_j \vee \Diamond A_j \in \varGamma$, where $\Box\neg A_j, \neg A_j \in \varDelta^{\boxdot}_{T}$. Then $C \rhd \bigvee \neg S_i \vee \bigvee A_j \in \varGamma$ and thus by P we obtain $C \rhd \bigvee \neg S_i \vee \bigvee A_j \in \varLambda$. Since $\varGamma \prec_S \varLambda$ we have $\Box \bigwedge S_i \in \varLambda$ so we obtain $C \rhd \bigvee A_j \in \varLambda$. But for each A_j we have $A_j \rhd \bigvee \neg T_{jk} \in \varLambda$ and thus $C \rhd \bigvee T_{jk} \in \varLambda$. Since $\varLambda \prec_T \varDelta$ we conclude $\neg C, \Box\neg C \in \varDelta$.

Lemma 19. *For logics containing* M$_0$ *we have* $\varGamma \prec_S \varDelta \prec \varDelta' \Rightarrow \varGamma \prec_{S \cup \varDelta^{\Box}_{\emptyset}} \varDelta'$.

Proof. Suppose $C \rhd \bigvee S_i \vee \bigvee \Diamond A_j \in \varGamma$, where $\Box\neg A_j \in \varDelta^{\Box}_{\emptyset}$. By M$_0$ we obtain $\Diamond C \wedge \bigwedge \Box\neg A_j \rhd \bigvee S_i \in \varGamma$. So, since $\varGamma \prec_S \varDelta$ and $\bigwedge \Box\neg A_j \in \varDelta$ we obtain $\Box\neg C \in \varDelta$ and thus $\Box\neg C, \neg C \in \varDelta'$.

Lemma 20. *For logics containing* R *we have* $\varGamma \prec_S \varDelta \prec_T \varDelta' \Rightarrow \varGamma \prec_{S \cup \varDelta^{\Box}_{T}} \varDelta'$.

[9] The reader should note that we do not give a completeness proof for ILR here. We only indicate a few problems one will encounter and indicate the usefulness of assuringness by overcoming these. In general assuringness does not yet give the answer to all problems encountered in modal completeness proofs. However, in the special case of ILRW assuringness can be put to use to give a completeness proof.

Proof. We consider A such that for some $S_i \in S$ and some $\Box\neg A_j \in \Delta^{\Box}_{T}$, we have $(A \rhd \bigvee \neg S_i \vee \bigvee \Diamond A_j) \in \Gamma$. By R we obtain $(\neg(A \rhd \bigvee A_j) \rhd \bigvee \neg S_i) \in \Gamma$, thus by $\Gamma {\prec_S} \Delta$ we get $(A \rhd \bigvee A_j) \in \Delta$. As $(A_j \rhd \bigvee \neg T_{kj}) \in \Delta$, also $(A \rhd \bigvee \neg T_{kj}) \in \Delta$. By $\Delta {\prec_T} \Delta'$ we conclude $\boxdot\neg A \in \Delta'$.

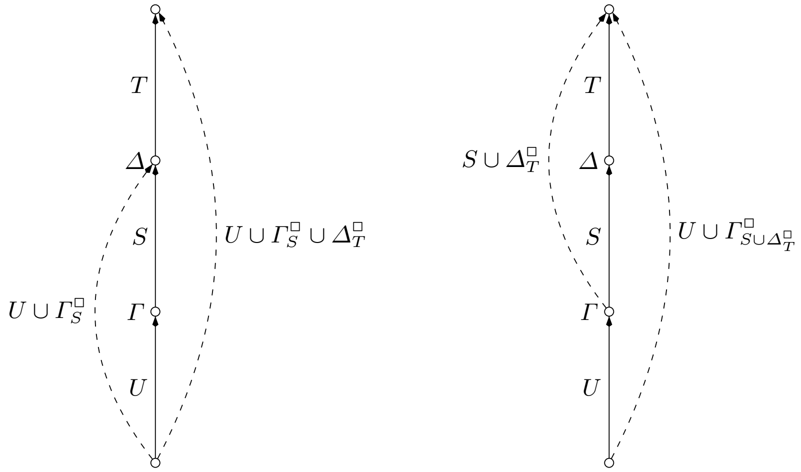
T
Δ
S
Γ
U
$U \cup \Gamma^{\Box}_{S}$
$U \cup \Gamma^{\Box}_{S} \cup \Delta^{\Box}_{T}$
T
$S \cup \Delta^{\Box}_{T}$
Δ
S
$U \cup \Gamma^{\Box}_{S \cup \Delta^{\Box}_{T}}$
Γ
U

Fig. 4. Two ways for computing the transitive closure in ILR.

What lemmata 18, 19 and 20 actually tell us is how to label R relations when we take R transitive while working in the lemma's respective logic. However, there is an easily identifiable problem here. Suppose we are working in ILR. Consider the two pictures in Figure 4. If we compute the label between the lower world and the upper world it does make a difference whether we first compute the label between the lower world and Δ (left picture) or the label between Γ and the upper world (right picture). We will show in Lemma 21 below that in the situation as given in Figure 4 we have

$$U \cup \Gamma^{\Box}_{S \cup \Delta^{\Box}_{T}} \subseteq U \cup \Gamma^{\Box}_{S} \cup \Delta^{\Box}_{T}.$$

And we should thus opt for the strategy as depicted in the leftmost picture when computing the transitive closure of R.

Lemma 21. *For logics containing*[10] R *we have* $\Gamma {\prec_S} \Delta \Rightarrow \Gamma^{\Box}_{S \cup \Delta^{\Box}_{T}} \subseteq \Delta^{\Box}_{T}$.

Proof. Consider $\Box\neg A \in \Gamma^{\Box}_{S \cup \Delta^{\Box}_{T}}$, that is, for some $S_i \in S$ and $\Box\neg B_j \in \Delta^{\Box}_{T}$, $A \rhd \bigvee \neg S_i \vee \bigvee \neg\Box\neg B_j \in \Gamma$. By R, $\neg(A \rhd \bigvee B_j) \rhd \bigvee \neg S_i \in \Gamma$, whence by $\Gamma {\prec_S} \Delta$, we get $A \rhd \bigvee B_j \in \Delta$. But for each B_j there is $T_{jk} \in T$ with $B_j \rhd \bigvee \neg T_{jk} \in \Delta$, whence $A \rhd \bigvee \neg T_{jk} \in \Delta$ and $\Box\neg A \in \Delta^{\Box}_{T}$.

[10] For the other logics we get similar lemmata.

Lemmata as Lemma 17, 19 and 20 are what we call *labelling lemma.* We propose the following slogan.

> **Slogan:** Every complete logic with a first order frame condition has its own labeling lemma.

Let us state two lemmata for ILW, a logic without a first order frame property. As predicted by our slogan, these do not fit in very nicely with the previous ones.

Lemma 22. *Suppose* $\neg(A \rhd B) \in \Gamma$. *There exists some* Δ *with* $\Gamma \prec_{\{\Box\neg A, \neg B\}} \Delta$ *and* $A \in \Delta$.

Proof. Suppose for a contradiction that there is no such Δ. Then there is a formula E with $(E \rhd \Diamond A \vee B) \in \Gamma$ such that $A, \neg E, \Box\neg E \vdash \bot$ and so $\vdash A \rhd E$. Then $(A \rhd \Diamond A \vee B) \in \Gamma$ and by the principle W we have $A \rhd B \in \Gamma$. The contradiction.

Lemma 23. *For logics containing* W *we have that if* $B \rhd C \in \Gamma \prec_S \Delta \ni B$ *then there exists* Δ *with* $\Gamma \prec_{S\cup\{\Box\neg B\}} \Delta \ni C, \Box\neg C$.

Proof. Suppose for a contradiction that no such Δ exists. Then for some formula A with $(A \rhd \bigvee \neg S_i \vee \Diamond B) \in \Gamma$, we get $C, \Box\neg C, \neg A, \Box\neg A \vdash \bot$, whence $\vdash C \rhd A$. Thus $B \rhd C \rhd A \rhd \bigvee \neg S_i \vee \Diamond B \in \Gamma$. By W, $B \rhd \bigvee \neg S_i \in \Gamma$ which contradicts $\Gamma \prec_S \Delta \ni B$.

4 Conclusion

In this paper two steps were taken.

The first one, presented in Section 2, involves the interpretability logic of all reasonable arithmentical theories. A new principle was formulated and the authors have the hope (and evidence) that proceeding along similar lines, much more can be achieved.

The second step involves a uniform treatment of different interpretability logics. Although much is know about the modal behavior of these logics, a relativelly clear and complete treatment (something that approaches a canonical model construction [1], Gentzen-style proof system [21]) is only given for the base logic IL. We hope that the machinery of Section 3 are the first steps to a uniform treatment of different modal interpretability logics.

References

1. de Jongh, D.H.J., Japaridze, G.K.: The Logic of Provability. In Buss, S.R., ed.: Handbook of Proof Theory. Studies in Logic and the Foundations of Mathematics, Vol.137. Elsevier, Amsterdam (1998) 475–546
2. Tarski, A., Mostowski, A., Robinson, R.: Undecidable theories. North–Holland, Amsterdam (1953)

3. Greenberg, M.: Euclidean and Non-Euclidean Geometries, 3d edition. Freeman (1996)
4. Gödel, K.: The Consistency of the Axiom of Choice and the Generalized Continuum Hypothesis with the Axioms of Set Theory. Volume 3 of Annals of Mathematical Studies. Princeton University Press (1940)
5. Gödel, K.: Über formal unentscheidbare Sätze der Principia Mathematica und verwnadter Systeme I. Monatsh. Math. Physik **38** (1931) 173–198
6. Simpson, S.G.: Partial realizations of Hilbert's program. Journal of Symbolic Logic **53** (1988) 349–363
7. Feferman, S.: Hilbert's program relativized: proof-theoretical and foundational reductions. Journal of Symbolic Logic **53** (1988) 364–384
8. Nelson, E.: Predicative arithmetic. Princeton University Press, Princeton (1986)
9. Shavrukov, V.Y.: The logic of relative interpretability over Peano arithmetic (in Russian). Technical Report 5, Stekhlov Mathematical Institute, Moscow (1988)
10. Berarducci, A.: The interpretability logic of Peano arithmetic. Journal of Symbolic Logic **55** (1990) 1059–1089
11. Visser, A.: Interpretability logic. In Petkov, P., ed.: Mathematical Logic, Proceedings of the 1988 Heyting Conference. Plenum Press, New York (1990) 175–210
12. Joosten, J.J.: Interpretability formalized. PhD thesis, Utrecht University (2004) ISBN: 90-393-3869-8.
13. Krajíček, J.: Bounded arithmetic, propositional logic, and complexity theory. Cambridge University Press, New York, NY, USA (1995)
14. Joosten, J.J., Goris, E.: Modal Matters in Interpretability Logics. Preprint LGPS-226, University of Utrecht (2004)
15. de Jongh, D.H.J., Veltman, F.: Modal completeness of ILW. In Gerbrandy, J., Marx, M., Rijke, M., Venema, Y., eds.: Essays dedicated to Johan van Benthem on the occasion of his 50th birthday. Amsterdam University Press, Amsterdam (1999)
16. Visser, A.: An overview of Interpretability Logic. In Kracht, M., Rijke, M., Wansing, H., eds.: Advances in modal logic '96. CSLI Publications, Stanford, CA (1997) 307–359
17. Joosten, J.: Towards the interpretability logic of all reasonable arithmetical theories. Master's thesis, University of Amsterdam (1998)
18. Joosten, J., Visser, A.: How to derive principles of interpretability logic, A toolkit. In: Liber Amicorum for Dick de Jongh. Intitute for Logic, Language and Computation (2004) Electronically published, ISBN: 90 5776 1289.
19. Joosten, J.J., Biklova, M., Goris, E.: Full labels. In: Liber Amicorum for Dick de Jongh. Intitute for Logic, Language and Computation (2004) Electronically published, ISBN: 90 5776 1289.
20. de Jongh, D.H.J., Veltman, F.: Provability logics for relative interpretability. In Petkov, P., ed.: Mathematical logic, Proceedings of the Heyting 1988 Summer School. Plenum Press (1990) 31–42
21. Sasaki, K.: A Cut-free Sequent System for the Smallest Interpretability Logic. Studia Logica **70** (2001) 353–372