The Interpolation Theorem for IL and ILP

Carlos Areces Dick de Jongh Eva Hoogland
ILLC, Universiteit van Amsterdam
Plantage Muidergracht 24, 1018 TV Amsterdam, The Netherlands
dickdj@wins.uva.nl

November 3, 1998

Abstract

In this article we establish interpolation for the minimal system of interpretability logic IL. We prove that arrow interpolation holds for IL and that turnstile interpolation and interpolation for the $\rhd$-modality easily follow from this. Furthermore, these properties are extended to the system ILP. The related issue of Beth Definability is also addressed. As usual, the arrow interpolation property implies the Beth property. From the latter it follows via an argumentation which is standard in provability logic, that IL has the fixed point property. Finally we observe that a general result of Maksimova [11] for provability logics can be extended to interpretability logics, implying that all extensions of IL have the Beth property.

Keywords Interpretability Logic, Interpolation Properties, Beth Property, Fixed Point Property.

1 Introduction

1.1 Some History

Interpretability logics are extensions of provability logics introduced by Visser in [15]. There the modal logics IL, ILM and ILP are defined by extending the object language of the basic provability logic L with a binary operator $\rhd$. This modality is to be read, relative to an (arithmetical) theory T, as: $A \rhd B$ iff $T + B$ is relatively interpretable in $T + A$. To put it simply, there is a function f (the interpretation) on the formulas of the language of T such that $T + B \vdash C \Rightarrow T + A \vdash f(C)$. (Obviously this translation should satisfy certain further requirements.) The main importance of interpretability logics is that they permit a finer analysis of arithmetical theories. For example, where the provability operators $\Box_{\mathsf{PA}}$ and $\Box_{\mathsf{GB}}$[1] have the same properties, the interpretability operator for PA and the one for GB differ: $\rhd_{\mathsf{PA}}$ satisfies the axiom $M : A \rhd B \to (A \wedge \Box C) \rhd (B \wedge \Box C)$, whereas $\rhd_{\mathsf{GB}}$ satisfies the axiom $P : A \rhd B \to \Box(A \rhd B)$.

Interpretability logics are useful and powerful tools for the study of the strength of different theories. However, in this work we are only interested in interpretability logics as *systems of (non standard) modal logic*. In the present article we establish a purely theoretical result

[1] where PA is Peano's formalization of Arithmetic and GB is Gödel - Bernay's formalization of Set Theory.

about the interpretability logics IL and ILP stating that these systems have the interpolation property. Hereto, a simple modal reading of $\rhd$ over Kripke models suffices.

1.2 Interpretability and Interpolation

When a new logic is defined, some questions immediately come to mind as a yardstick by which to measure the behaviour of the newborn logic. Is the logic sound, complete, decidable? In this article we deal with one of these metalogical questions: Craig interpolation.

In [3] Craig proved his famous interpolation theorem for first-order logic (FO): Whenever $\vdash_{\mathsf{FO}} A \to B$, then there exists a formula I (the interpolant) in the common language of A and B such that $\vdash_{\mathsf{FO}} A \to I$ and $\vdash_{\mathsf{FO}} I \to B$. The interpolation property is a sign of a well-behaved deduction system. Besides its theoretical interest, this property plays a crucial role in, for example, the field of automated theorem proving, where it can be used to restrict the search space of the inference algorithm, in looking for intermediate lemmas.

For interpretability logics some (positive and negative) results about interpolation are known. In [16] a proof by Ignatiev of failure of interpolation for ILM is adapted, showing that systems between ILM_0 and ILM do not have interpolation (for the definition of the systems mentioned we refer to [16]). It follows for example that ILW^* does not have interpolation. In [4] de Rijke studies unary interpretability logic, i.e., the logic of $(\top \rhd \psi)$. He shows that the restricted systems il, ilp and ilm, all satisfy interpolation.

The question of interpolation for the basic system IL was raised by Baaz. Hájek [5] gave a positive answer to this question, but unfortunately overlooked some cases as was pointed out by Ignatiev. The latter fixed some of the cases in [7], but the proof remained incomplete for years. In this article we provide a full proof. The techniques developed for this proof also serve to establish interpolation for the system ILP. An alternative way of settling this question was given by Hájek who showed interpolation for ILP assuming that this property holds for IL (cf. [5]).

The following table summarizes the results in the field after our contribution.

Binary Systems	IL	ILP	ILM	ILF	ILW	ILW*
Interpolation	yes	yes	no	open	open	no
Proved in	This paper	[5] This paper	[6]			[16]
Unary Systems	**il**	**ilp**	**ilm**			
Interpolation	yes	yes	yes			
Proved in	[4]	[4]	[4]			

In this article we assume the reader is familiar with basic notions of modal logic in general, but we develop in detail the necessary concepts specifically devised in the context of provability and interpretability logics (Section 2). For a thorough introduction to the topic covering the arithmetical interest of this project we refer to [8] and [16]. Section 3 contains the main result of the present paper showing that arrow interpolation holds for IL. As corollaries we obtain in Section 4 that turnstile interpolation and interpolation for $\rhd$ also hold for IL. We also show that all these properties transfer to ILP. In the final section we comment on an interesting interplay between Beth Definability and Fixed Points in interpretability logics: As in provability logics, the Beth Theorem (derivable in a standard manner from arrow interpolation) can be used to give an alternative proof for the Fixed Point Theorem. We then extend Maksimova's general result concerning the Beth property for provability logics [11] to interpretability logics, implying that all extensions of IL have this property.

2 Preliminaries

We now gather some definitions and preliminary results needed for our main theorem. We start by defining the basic system of interpretability logic IL.

Definition 2.1 (The System IL) The *basic system for interpretability logic* IL is defined by the following axiom schemes:

$L1$ All classical tautologies.
$L2$ $\Box(A \to B) \to (\Box A \to \Box B)$.
$L3$ $\Box A \to \Box\Box A$.
$L4$ $\Box(\Box A \to A) \to \Box A$.
$J1$ $\Box(A \to B) \to A \rhd B$.
$J2$ $(A \rhd B \wedge B \rhd C) \to A \rhd C$.
$J3$ $(A \rhd C \wedge B \rhd C) \to (A \vee B) \rhd C$.
$J4$ $A \rhd B \to (\Diamond A \to \Diamond B)$.
$J5$ $\Diamond A \rhd A$.

together with the rules of Modus Ponens and Necessitation (i.e., $\vdash A \Rightarrow \vdash \Box A$). The notions of proof in IL and of theorems and rules are defined as usual. ◀

For some intuitions about the role of the above axioms let us turn for a moment to their arithmetical interpretation. Axioms $L1$ to $L4$ are the principles of Löb's Logic L, the basic system of provability logic; $J1$ says that the identity is an interpretation; $J2$ expresses transitivity of the $\rhd$-modality, reflecting that interpretations can be composed. By $J3$ two different interpretations can be joined in a definition by cases; $J4$ states that relative interpretability implies relative consistency; $J5$ is the 'Interpretation Existence Lemma' (cf. [16]), a formalization in arithmetic of Henkin's completeness theorem.

In the proof of Theorem 1 the following facts will be useful.

Proposition 2.2 ([8], [16]) *In IL the following theorems are derivable:*
1. $\vdash \Box D \leftrightarrow \neg D \rhd \bot$.
2. $\vdash (D \vee \Diamond D) \rhd D$.
3. $\vdash D \rhd (D \wedge \Box\neg D)$.
4. $\vdash ((D \wedge E) \rhd F) \to (\neg D \rhd F \to E \rhd F)$.

Proof of Proposition 2.2: Part (1),(2) and (4) are easy; (3) follows from the fact that in Löb's Logic L we can derive $\vdash_{\mathsf{L}} \Diamond D \to \Diamond(D \wedge \Box\neg D)$, and hence $\vdash_{\mathsf{L}} D \to (D \wedge \Box\neg D) \vee \Diamond(D \wedge \Box\neg D)$. Now apply (2). ■

We now turn to semantics. A Kripke semantics (in this case also called *Veltman semantics*) for IL was first presented in [9].

Definition 2.3 (IL-Frames, IL-Models, Forcing Relation) An IL-*frame* is a tuple $\langle W, R, S\rangle$, where:

- W is a non-empty set.
- R is a transitive, upwards well-founded binary relation on W.
- For each $w \in W$,

- S_w is a binary relation defined on $w\uparrow \stackrel{\text{def}}{=} \{u \in W : wRu\}$.
- S_w is transitive and reflexive.
- $wRuRv \Rightarrow uS_wv$.

An IL-*model* is a structure $\langle\langle W, R, S\rangle, V\rangle$, where $\langle W, R, S\rangle$ is an IL-frame and V is a modal valuation assigning subsets of W to proposition letters.
A *forcing relation* $\models$ on an IL-model satisfies the usual clauses for atomic formulas, Boolean connectives and $\Box$-modality (with R as the accessibility relation), plus the following extra clause:

- $w \models A \rhd B \Leftrightarrow \forall u((wRu \wedge u \models A) \rightarrow \exists v(uS_wv \wedge v \models B))$. ◀

De Jongh and Veltman [9] provide a *modal completeness theorem* for IL with respect to finite IL-models.

Note that the clause for the $\rhd$-modality in the definition of the forcing relation above, is unlike the clause for the usual $\Box$-modality. This is why we consider interpretability logics to be *non-standard* systems of modal logic.

Convention 2.4 In the rest of the section we will tacitly assume that we are working in IL. Hence all the notions defined below are to be read relative to this system. For example, when we speak about a set of formulas it will be understood that these are IL-formulas, etc.

The method we will use for showing interpolation will be a standard model-theoretic Henkin style proof as can be found, e.g., in the proof of interpolation for provability logic in [13]. The aim of these proofs is to construct a model of the logic under consideration whose worlds are based on maximal consistent sets of formulas. However, since IL is not compact, maximal consistent sets should be confined to finite adequate subsets of the language. Our first task is to specify this notion of adequateness.

Definition 2.5 ($\sim A$, Adequate Set, [9]) If the formula A is not a negation, then $\sim A$ is $\neg A$. Otherwise, if A is $\neg B$, then $\sim A$ is B. A set X of formulas is called *adequate* if X is closed under subformulas and the $\sim$-operation, $\bot \rhd \bot \in X$ and X contains $A \rhd B$ whenever A, B are antecedent or succedent of a $\rhd$-formula in X. ◀

From this point onwards it is best to consider $\Box A$ as an abbreviation of $\sim A \rhd \bot$. This is allowed by Proposition 2.2.1. In particular, this implies that whenever formulas of the form $\Box\neg A, \Box\neg B$ are contained in an adequate set X, then also $A \rhd B \in X$.

Notation 2.6 For any set of formulas X there exists a smallest adequate set containing X, denoted by $\mathcal{A}_X$. As usual, we will omit brackets. By $\mathcal{L}_X$ (read: *the language of* X) we denote the set of IL-formulas built up from proposition letters occurring in formulas in X. For X a finite set of formulas, we interchangeably write X for its conjunction: e.g. $\vdash \bigwedge X \rightarrow A$ will be written simply as $\vdash X \rightarrow A$. ◀

Remark 2.7 Note that if X is finite, then so is $\mathcal{A}_X$, as desired. In order to ensure this, the set X in Definition 2.5 was required to be closed under negation of non-negated formulas only. ◀

In modal logic, proofs of interpolation are in general close in spirit to completeness proofs. The centrale role played by *maximal consistent sets* in the latter is in the former taken over by *complete inseperable pairs.*

Definition 2.8 (Inseparable Pair) A pair $\langle X, Y\rangle$ of finite sets of formulas is called *separable* if there exists a formula $A \in \mathcal{L}_X \cap \mathcal{L}_Y$ such that $\vdash X \to A$ and $\vdash Y \to \neg A$. A pair is called *inseparable* if it is not separable. ◀

Note that for any inseparable pair $\langle X, Y\rangle$, the sets X and Y are each consistent.

Definition 2.9 (Complete Pair) Let $\langle X, Y\rangle$ be an inseparable pair. We say that $\langle X, Y\rangle$ is *complete* if

1. For each $A \in \mathcal{A}_X$, either $A \in X$ or $\sim A \in X$.
2. For each $A \in \mathcal{A}_Y$, either $A \in Y$ or $\sim A \in Y$. ◀

In e.g. [14] the following analogue of Lindenbaum's Lemma can be found.

Proposition 2.10 *Let $\langle X, Y\rangle$ be an inseparable pair. Then there exist sets X', Y' such that $X \subseteq X' \subseteq \mathcal{A}_X$, $Y \subseteq Y' \subseteq \mathcal{A}_Y$ and $\langle X', Y'\rangle$ is a complete pair.*

The preparations up to now suffice to define the worlds of the construction we are after. To define the relations in this model the following notion is needed.

Definition 2.11 ($\prec$ Relation) Let $\langle X, Y\rangle$, $\langle X', Y'\rangle$ be two complete pairs such that $\mathcal{A}_X = \mathcal{A}_{X'}$, $\mathcal{A}_Y = \mathcal{A}_{Y'}$. We put $\langle X, Y\rangle \prec \langle X', Y'\rangle$ if

1. For each A, if $\Box A \in X \cup Y$ then $\Box A, A \in X' \cup Y'$.
2. There exists some A such that $\Box A \notin X \cup Y$ but $\Box A \in X' \cup Y'$. ◀

The above is the canonical definition of the accessibility relation for the $\Box$-modality which takes care of the conditions of transitivity and upward well-foundedness.

In order to motivate the next definition, let us jump a little bit ahead of ourselves, and ask what this entire enterprise should amount to. As usual in Henkin-style proofs for interpolation, the idea is the following. On the assumption that some two formulas B and C (such that $\vdash B \to C$) *do not* have an interpolant, the pair $\langle\{B\}, \{\neg C\}\rangle$ can be extended to a complete pair which will be a world in the model that is now to be constructed. The key point is then to prove a truth lemma for the eventual model saying that a formula is valid in a world if and only if that formula is contained in one component of the complete pair which constitutes that world. This lemma implies that we have constructed a world in which B and $\neg C$ holds, contrary to the fact that $B \to C$ is a theorem and we are done. Now, for proving the truth lemma we will in particular have to show that, if a formula of the form $\neg(G \rhd A)$ is contained in some world w, then $w \not\models (G \rhd A)$. According to the truth definition, we should in that case produce an R-successor u of w which contains G and which 'avoids' A in the sense that any S_w-successor of u does not contain A.

What makes this concept of "A-avoiding" hard to grasp, is the fact that avoiding a formula A involves other formulas D as well. Let us see why. Hereto, consider a world w which contains a formula of the form $D \rhd A$. In this case any truly A-avoiding successor u of w

is not allowed to contain D, nor to have an R-successor v containing D. Else, by the truth lemma, $w \models D \rhd A$. In the first case it follows directly from the truth definition that u has an S_w-successor satisfying A, contrary to u being A-avoiding. In the second case we reason as follows. Since wRv (by transitivity of R) it follows again from the truth-definition that v has an S_w-successor z which contains A. Moreover, $wRuRv$ and hence, by the definition of IL-frame, uS_wv. Since S_w is transitive, this shows that z is a S_w-successor of u, and again we end up with an S_w-successor of u containing A. Bearing this in mind, a first attempt to formulate the notion of 'A-avoiding successor' would be the following.

Definition 2.12 (A-Critical, preliminary, [5]) Let $\langle X, Y\rangle$, $\langle X', Y'\rangle$ be two complete pairs such that $\mathcal{A}_X = \mathcal{A}_{X'}$, $\mathcal{A}_Y = \mathcal{A}_{Y'}$. Let $\Box\neg A \in \mathcal{A}_X \cup \mathcal{A}_Y$. We say that $\langle X', Y'\rangle$ is an *A-critical successor* of $\langle X, Y\rangle$ if the following conditions are met.

1. $\langle X, Y\rangle \prec \langle X', Y'\rangle$.
2. $X_1 \stackrel{\text{def}}{=} \{\neg D, \Box\neg D : D \rhd A \in X\} \subseteq X'$.
   $Y_1 \stackrel{\text{def}}{=} \{\neg E, \Box\neg E : E \rhd A \in Y\} \subseteq Y'$. ◄

However complicated as the above definition may seem, it does not yet suffice since it does not reckon with a possible interplay between formulas from $\mathcal{A}_X$ and $\mathcal{A}_Y$. To make this point more precise, let us imagine the situation where $A \in \mathcal{A}_X \setminus \mathcal{A}_Y$ and $B \in \mathcal{A}_Y \setminus \mathcal{A}_X$. Although the formulas A and B come from entirely different adequate sets, still B can turn out to be an undesirable member of any A-critical successor of a pair $\langle X, Y\rangle$. For it can be the case that $\vdash X \to C \rhd A$ and $\vdash Y \to B \rhd C$, for some $C \in \mathcal{L}_X \cap \mathcal{L}_Y$ but *not necessarily in $\mathcal{A}_X$ or $\mathcal{A}_Y$*. By soundness then $\langle X, Y\rangle \models B \rhd A$, and B should henceforth be avoided as not to run in the same trouble as before. However, since $B \rhd A$ is not contained in any of the adequate sets $\mathcal{A}_X, \mathcal{A}_Y$, and hence $B \rhd A \notin X \cup Y$, Definition 2.12 does not give any restrictions in this case. On these grounds we exchange our preliminary definition for the one below.

Definition 2.13 (A-Critical) Let $\langle X, Y\rangle$, $\langle X', Y'\rangle$ be two complete pairs such that $\mathcal{A}_X = \mathcal{A}_{X'}$, $\mathcal{A}_Y = \mathcal{A}_{Y'}$. Let $\Box\neg A \in \mathcal{A}_X \cup \mathcal{A}_Y$. We say that $\langle X', Y'\rangle$ is an *A-critical successor of* $\langle X, Y\rangle$ (notation: $\langle X, Y\rangle \prec_A \langle X', Y'\rangle$), if the following conditions are met.

1. $\langle X, Y\rangle \prec \langle X', Y'\rangle$.
2. If $\Box\neg A \in \mathcal{A}_X$, then
   $X_1 \stackrel{\text{def}}{=} \{\neg D, \Box\neg D : D \rhd A \in X\} \subseteq X'$.
   $Y_1 \stackrel{\text{def}}{=} \{\neg E, \Box\neg E : \Box\neg E \in \mathcal{A}_Y \ \&\ \exists C \in \mathcal{L}_X \cap \mathcal{L}_Y [\vdash Y \to (E \rhd C) \ \&\ \vdash X \to (C \rhd A)]\} \subseteq Y'$.
3. If $\Box\neg A \in \mathcal{A}_Y$, then
   $X_2 \stackrel{\text{def}}{=} \{\neg D, \Box\neg D : \Box\neg D \in \mathcal{A}_X \ \&\ \exists C \in \mathcal{L}_X \cap \mathcal{L}_Y [\vdash X \to (D \rhd C) \ \&\ \vdash Y \to (C \rhd A)]\} \subseteq X'$.
   $Y_2 \stackrel{\text{def}}{=} \{\neg E, \Box\neg E : E \rhd A \in Y\} \subseteq Y'$. ◄

Note that the complications described above only occur in case A and B are contained in different adequate sets. That is why the sets X_1 and Y_2 in Definition 2.13 remain unaltered as compared to the sets X_1, Y_1 in Definition 2.12.

Summarizing, the difficulties in finding the above notion of criticality which will turn out to be the one needed for the interpolation proof were twofold. First, the non-standard character of the $\rhd$-modality brought on the problem that avoiding one formula involves other formulas. Second, the fact that we are interested in interpolation made us pay attention to the languages. The next claim implies that the above notion is well-defined.

Claim 2.14 *If $\Box\neg A \in \mathcal{A}_X \cap \mathcal{A}_Y$ in Definition 2.13, then $X_1 = X_2$ and $Y_1 = Y_2$.*

Proof of Claim 2.14: Let $\Box\neg A \in \mathcal{A}_X \cap \mathcal{A}_Y$. Obviously $X_1 \subseteq X_2$. For the other inclusion, consider a formula D such that $\neg D, \Box\neg D \in X_2$. That is, $\Box\neg D \in \mathcal{A}_X$ and there exists some $C \in \mathcal{L}_X \cap \mathcal{L}_Y$ such that $\vdash X \to (D \rhd C)$ (*) and $\vdash Y \to (C \rhd A)$. We want to show that $\neg D, \Box\neg D \in X_1$, i.e., $D \rhd A \in X$. Let us assume for contradiction that $D \rhd A \notin X$. Since $D \rhd A \in \mathcal{A}_X$, by completeness of $\langle X, Y\rangle$ this assumption implies that $\neg(D \rhd A) \in X$ (**). By (*), $\vdash X \to [(C \rhd A) \to (D \rhd A)]$. From (**) it now follows that $\vdash X \to \neg(C \rhd A)$. We conclude that $C \rhd A$ separates X and Y. Contradiction. To show that $Y_1 = Y_2$, one proceeds analogously. ■

Note that for any $\langle X, Y\rangle$, $\langle X', Y'\rangle$, $\langle X'', Y''\rangle$ and any formula A we have that

$$\langle X, Y\rangle \prec_A \langle X', Y'\rangle \prec \langle X'', Y''\rangle \Longrightarrow \langle X, Y\rangle \prec_A \langle X'', Y''\rangle.$$

This finishes the necessary preliminaries for the next section.

3 The Interpolation Theorem for IL

The next theorem is the main result of this paper.

THEOREM 1 (The Arrow Interpolation Theorem for IL) *Let D_0, E_0 be IL-formulas. Assume $\vdash_{\mathsf{IL}} D_0 \to E_0$. Then there exists an IL-formula $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that $\vdash_{\mathsf{IL}} D_0 \to I$ and $\vdash_{\mathsf{IL}} I \to E_0$.*

Proof of Theorem 1: Let $\vdash_{\mathsf{IL}} D_0 \to E_0$. Assume there is no interpolant. In the next few pages it will be shown that this assumption enables us to construct an IL-model which contains a world satisfying both D_0 and $\neg E_0$. From the soundness of IL a contradiction follows. Now let us get to work.

By assumption $\vdash_{\mathsf{IL}} D_0 \to E_0$ has no interpolant. In other words, $\langle \{D_0\}, \{\neg E_0\}\rangle$ is inseparable. By Proposition 2.10 there exist sets X_0, Y_0 such that $\{D_0\} \subseteq X_0 \subseteq \mathcal{A}_{D_0}$, $\{\neg E_0\} \subseteq Y_0 \subseteq \mathcal{A}_{E_0}$ and $\langle X_0, Y_0\rangle$ is a complete pair. We define the model $\mathcal{M} \stackrel{\text{def}}{=} \langle\langle W, R, S\rangle, V\rangle$ as follows.

Begin construction of model.

- Each world in W will be a sequence of 2-tuples consisting of a complete pair together with a sequence of formulas recording 'how we arrived at that pair'. Let [] represent the empty sequence and $*$ stand for concatenation. Formally, W is the smallest set satisfying the following two conditions:
  - $w_0 \stackrel{\text{def}}{=} [(\langle X_0, Y_0\rangle, [])] \in W$.
  - Let $[(\langle X_0, Y_0\rangle, []), \ldots, (\langle X_n, Y_n\rangle, \tau_n)] \in W$. Let $\langle X, Y\rangle$ be a complete pair such that $X \subseteq \mathcal{A}_{D_0}$ and $Y \subseteq \mathcal{A}_{E_0}$, and moreover $\langle X_n, Y_n\rangle \prec_A \langle X, Y\rangle$ for some A. Then $[(\langle X_0, Y_0\rangle, []), \ldots, (\langle X_n, Y_n\rangle, \tau_n), (\langle X, Y\rangle, \tau_n * [A])] \in W$.

Notation 3.1 For all $w \in W$, $w = [(\langle X_0, Y_0\rangle, []), \ldots, (\langle X_n, Y_n\rangle, \tau_n)]$ we will write X_w (resp. Y_w, τ_w) for the set X_n (resp. Y_n, τ_n). For $w, u \in W$, the notation $w \subseteq u$ (resp. $w \subset u$) indicates that w is an initial (resp. *proper* initial) segment of u. ◀

- For all $w, u \in W$, we define wRu iff $w \subset u$.
- For all $w, u, v \in W$, we define uS_wv iff there exists some formula A and complete pairs $\langle X', Y' \rangle$, $\langle X'', Y'' \rangle$ such that $w * [(\langle X', Y' \rangle, \tau_w * [A])] \subseteq u$, $w * [(\langle X'', Y'' \rangle, \tau_w * [A])] \subseteq v$.

We leave it to the reader to check that $\langle W, R, S \rangle$ is an IL-frame. That is, W is finite, R is transitive and irreflexive, and S_w is a transitive and reflexive relation defined over the set $\{u \in W : wRu\}$ such that for every $w', w'' \in W$ we have that $wRw'Rw''$ implies $w'S_ww''$.

Finally, for every $w \in W$ and every proposition letter $p \in \mathcal{L}_{D_0} \cup \mathcal{L}_{E_0}$, we set the valuation V to

$$w \in V(p) \overset{\text{def}}{\Longleftrightarrow} p \in X_w \cup Y_w.$$

End of construction.

The proof of Theorem 1 now reduces to the following truth lemma.

Lemma 3.2 (Truth Lemma) *Let $\mathcal{M} = \langle \langle W, R, S \rangle, V \rangle$ be the model defined above. Then for any $w \in W$,*

1. *$B \in \mathcal{A}_{D_0}$ implies $w \models B \Leftrightarrow B \in X_w$, and*
2. *$B \in \mathcal{A}_{E_0}$ implies $w \models B \Leftrightarrow B \in Y_w$.*

Note that this in particular implies that $w_0 \models D_0$ and $w_0 \models \neg E_0$, for $w_0 \in W$ defined above. Hence this lemma is all that stands between us and a proof of Theorem 1.

The hard part of proving the Truth Lemma is summarized in the two lemmas below, the proof of which is postponed till their use has been demonstrated.

Notation 3.3 For all $w, u \in W$, and any formula A,

$$wR_Au \overset{\text{def}}{\Longleftrightarrow} \text{there exists } \langle X', Y' \rangle \text{ such that } w * [(\langle X', Y' \rangle, \tau_w * [A])] \subseteq u. \qquad \blacktriangleleft$$

So, wR_Au implies that $\langle X_w, Y_w \rangle \prec_A \langle X_u, Y_u \rangle$.

Lemma 3.4 *Let $\neg(G \rhd F) \in X_w$ (resp. Y_w). Then there exists some $u \in W$ such that wR_Fu and $G \in X_u$ (resp. Y_u).*

Lemma 3.5 *Let $G \rhd F \in X_w$ (resp. Y_w). Let $u \in W$ be such that wR_Au and $G \in X_u$ (resp. Y_u). Then there exists $v \in W$ such that wR_Av and $F \in X_v$ (resp. Y_v).*

Proof of Truth Lemma: This proof is by induction on the complexity of B. The atomic case is given by definition, the Boolean cases are an easy exercise and the $\Box$-case is an instance of the $\rhd$-case. Hence let us concentrate on the latter.

Let B be of the form $G \rhd F \in \mathcal{A}_{D_0} \cup \mathcal{A}_{E_0}$. Let us assume that $G \rhd F \in \mathcal{A}_{D_0}$ (in case that $(G \rhd F) \in \mathcal{A}_{E_0}$ we reason similarly).

CASE "$\Rightarrow$" : Let $G \rhd F \notin X_w$. By completeness of $\langle X_w, Y_w \rangle$, then $\neg(G \rhd F) \in X_w$. By Lemma 3.4, no S_w-successor v of the element u produced there, satisfies F: for, wR_Fu and uS_wv imply that wR_Fv. Since $F \rhd F \in X_w$, it follows that $v \not\models F$. We conclude that $w \not\models G \rhd F$, and we are done.

CASE "$\Leftarrow$" : Let $G \rhd F \in X_w$. Let $u \in W$ be such that wRu and $u \models G$. Then $wR_A u$, for some formula A. By induction hypothesis, $G \in X_u$. By Lemma 3.5 there exists some $v \in W$ such that $uS_w v$ and $F \in X_v$. Again by the induction hypothesis $v \models F$, and it follows that $w \models G \rhd F$.

Q.E.D. Truth lemma.

Now let us prove the two auxiliary lemmas. Both lemmas will be shown to hold for X_w, X_u, X_v. For Y_w, Y_u, Y_v, the proofs are similar.

Proof of Lemma 3.4: Let $\neg(G \rhd F) \in X_w$. We define

$$X^- \stackrel{\text{def}}{=} \bigodot X_w \cup \{G, \Box\neg G\} \cup \{\neg D, \Box\neg D : D \rhd F \in X_w\},$$

$$Y^- \stackrel{\text{def}}{=} \bigodot Y_w \cup \{\neg E, \Box\neg E : \Box\neg E \in \mathcal{A}_{E_0} \ \& \ \exists C \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0} [\vdash Y_w \to (E \rhd C) \ \& \ \vdash X_w \to (C \rhd F)]\},$$

where here, as elsewhere in the proof, for any set of formulas X,

$$\bigodot X \stackrel{\text{def}}{=} \{D, \Box D : \Box D \in X\}.$$

We will show that X^- and Y^- are inseparable. For then, by Proposition 2.10 we can extend $\langle X^-, Y^- \rangle$ to a complete pair $\langle X_u, Y_u \rangle$, and the element $u \stackrel{\text{def}}{=} w * [(\langle X_u, Y_u \rangle, \tau_w * [F])]$ will satisfy all our requirements.

Let us assume for contradiction that X^- and Y^- are separable. That is, there exists some $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that

$$\vdash X^- \to I \qquad \text{and} \qquad \vdash Y^- \to \neg I.$$

Now we can derive the following:

$$\vdash \bigodot X_w \to [(G \wedge \Box\neg G \wedge \neg I) \to \bigvee_{D \rhd F \in X_w} (D \vee \Diamond D)].$$

Henceforth we will simply omit the indexset (in this case X_w) over which a disjunction is taken, in case this set is clear from the context. Reasoning as in provability logic, we obtain from the definition of $\bigodot X_w$ and axiom $J1$ that

$$\vdash X_w \to [(G \wedge \Box\neg G \wedge \neg I) \rhd \bigvee (D \vee \Diamond D)].$$

By Proposition 2.2.2 and the fact that $D \rhd F \in X_w$, then

$$\vdash X_w \to [(G \wedge \Box\neg G \wedge \neg I) \rhd F].$$

With the help of Proposition 2.2.4 we derive that

$$\vdash X_w \to [(I \rhd F) \to ((G \wedge \Box\neg G) \rhd F)]. \tag{1}$$

On the other hand,

$$\vdash \bigodot Y_w \to [I \to \bigvee (E_j \vee \Diamond E_j)],$$

for some finite index set J. The formulas E_j are such that there exist $C_j \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ for which

$$\vdash Y_w \to [I \rhd (\bigvee C_j)] \qquad \text{and} \qquad \vdash X_w \to [(\bigvee C_j) \rhd F] \qquad \text{holds.}$$

It follows that

$$\vdash X_w \to [(I \rhd (\bigvee C_j)) \to (I \rhd F)].$$

Together with (1) and the fact that $\neg(G \rhd F) \in X_w$ this implies via Proposition 2.2.3 that

$$\vdash X_w \to [\neg(I \rhd (\bigvee C_j))].$$

Hence $I \rhd (\bigvee C_j)$ separates X_w and Y_w. A contradiction.
Q.E.D. Lemma 3.4.

Proof of Lemma 3.5: Let $G \rhd F \in X_w$. Let $u \in W$ be such that $wR_A u$ and $G \in X_u$. By definition of criticality, $\Box\neg A \in \mathcal{A}_{D_0} \cup \mathcal{A}_{E_0}$. In this proof we distinguish as to whether $\Box\neg A \in \mathcal{A}_{D_0}$ or $\Box\neg A \in \mathcal{A}_{E_0}$.

CASE 1: Let $\Box\neg A \in \mathcal{A}_{D_0}$. Analogously to the proof of Lemma 3.4 we define

$$X^- \stackrel{\text{def}}{=} \bigodot X_w \cup \{F, \Box\neg F\} \cup \{\neg D, \Box\neg D : D \rhd A \in X_w\},$$
$$Y^- \stackrel{\text{def}}{=} \bigodot Y_w \cup \{\neg E, \Box\neg E : \Box\neg E \in \mathcal{A}_{E_0} \;\&\; \exists C \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0} [\vdash Y_w \to (E \rhd C) \;\&\; \vdash X_w \to (C \rhd A)]\}.$$

Again we will show that $\langle X^-, Y^- \rangle$ can be extended to a complete pair $\langle X_v, Y_v \rangle$. Then, the element $v \stackrel{\text{def}}{=} w * [(\langle X_v, Y_v \rangle, \tau_w * [A])]$ will have all the required properties.

So, let us assume for contradiction that there exists some $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that

$$\vdash X^- \to I \qquad \text{and} \qquad \vdash Y^- \to \neg I.$$

Again we derive that

$$\vdash \bigodot X_w \to [(F \wedge \Box\neg F \wedge \neg I) \to \bigvee (D \vee \Diamond D)].$$

Reasoning as before we see that

$$\vdash X_w \to [(F \wedge \Box\neg F \wedge \neg I) \rhd A],$$

and

$$\vdash X_w \to [(I \rhd A) \to (F \wedge \Box\neg F \rhd A)]. \tag{2}$$

Since $G \rhd F \in X_w$ one immediately sees that

$$\vdash X_w \to [(F \rhd A) \to (G \rhd A)]. \tag{3}$$

Now assume that $(G \rhd A) \in X_w$. Since $wR_A u$, then $\neg G \in X_u$, which by assumption is not the case. We conclude that $(G \rhd A) \notin X_w$, hence by completeness of $\langle X_w, Y_w \rangle$

$$\neg(G \rhd A) \in X_w. \tag{4}$$

On the other hand,

$$\vdash \bigodot Y_w \to [I \to \bigvee (E_j \vee \Diamond E_j)],$$

for some finite index set J. The formulas E_j are chosen in such a way that there exist formulas $C_j \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that

$$\vdash Y_w \to [I \rhd (\bigvee C_j)] \qquad \text{and} \qquad \vdash X_w \to [(\bigvee C_j) \rhd A].$$

It follows that

$$\vdash X_w \to [(I \rhd (\bigvee C_j)) \to (I \rhd A)]. \tag{5}$$

(2), (3), (4), (5) and Proposition 2.2.3 together imply that

$$\vdash X_w \to [\neg(I \rhd (\bigvee C_j))].$$

This shows that $(I \rhd (\bigvee C_j))$ separates X_w and Y_w, which is again a contradiction.

CASE 2: Let $\Box\neg A \in \mathcal{A}_{E_0}$. This time we define

$$X^- \stackrel{\text{def}}{=} \bigodot X_w \cup \{F, \Box\neg F\} \cup \{\neg D, \Box\neg D : \Box\neg D \in \mathcal{A}_{D_0} \;\&\; \exists C \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0} [\vdash X_w \to (D \rhd C) \;\&\; \vdash Y_w \to (C \rhd A)]\},$$

$$Y^- \stackrel{\text{def}}{=} \bigodot Y_w \cup \{\neg E, \Box\neg E : E \rhd A \in Y_w\}.$$

Again we assume for contradiction that there exists some $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that

$$\vdash X^- \to I \qquad \text{and} \qquad \vdash Y^- \to \neg I.$$

Now we reason as follows. First note that

$$\vdash \bigodot Y_w \to [I \to \bigvee(E \vee \Diamond E)],$$

where for every E it is the case that $(E \rhd A) \in Y_w$. Hence

$$\vdash Y_w \to [I \rhd A]. \tag{6}$$

Also,

$$\vdash \bigodot X_w \to [(F \wedge \Box\neg F) \to (I \vee \bigvee(D_j \vee \Diamond D_j))],$$

for some finite index set J. Since $G \rhd F \in X_w$ this implies by Proposition 2.2.3 that

$$\vdash X_w \to [G \rhd (I \vee \bigvee(D_j \vee \Diamond D_i))]. \tag{7}$$

The formulas D_j are such that there exist formulas $C_j \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ for which

$$\vdash Y_w \to [(\bigvee C_j) \rhd A], \qquad \text{and} \tag{8}$$

$$\vdash X_w \to [(\bigvee D_j) \rhd (\bigvee C_j)].$$

Then also $\vdash X_w \to [(I \vee (\bigvee D_j)) \rhd (I \vee (\bigvee C_j))]$, hence by (7),

$$\vdash X_w \to [G \rhd (I \vee (\bigvee C_j))]. \tag{9}$$

From (8) and (6) it follows that

$$\vdash Y_w \to [(I \vee (\bigvee C_j)) \rhd A]. \tag{10}$$

By definition of A-criticality, (9) and (10) imply that $\neg G \in X_{w'}$, for every A-critical successor w' of w. But $wR_A u$, and $G \in X_u$. Contradiction. Again we conclude that the pair $\langle X^-, Y^- \rangle$ is inseparable, and we can extend it to a complete pair $\langle X_v, Y_v \rangle$. The element $v \stackrel{\text{def}}{=} w * [(\langle X_v, Y_v \rangle, \tau_w * [A])]$ has all the required properties.

Q.E.D. Lemma 3.5.

Q.E.D. Theorem 1.

4 Derived Results on Interpolation

4.1 Different Interpolation Properties for IL

In the literature on interpolation we will find that this property is presented in many (in principle different) ways, depending on e.g. the consequence relation under consideration, or our understanding of a 'common' language. Perhaps the two best known definitions in this genre are the *arrow* interpolation considered so far and the *turnstile* interpolation (where $\to$ is replaced by $\vdash$), and their corresponding semantic versions. There is no general connection between these properties. However, via a Deduction Theorem one easily derives turnstile interpolation from arrow interpolation.

Proposition 4.1 (Deduction Theorem for IL) *For any pair of IL-formulas A and B, $A \vdash_{\mathsf{IL}} B$ iff $\vdash_{\mathsf{IL}} (A \wedge \Box A) \to B$.*

As a consequence of the above proposition and Theorem 1, IL also has turnstile interpolation.

Corollary 4.2 (Turnstile Interpolation for IL) *Let D_0, E_0 be IL-formulas. Assume $D_0 \vdash_{\mathsf{IL}} E_0$. Then there exists an IL-formula $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that $D_0 \vdash_{\mathsf{IL}} I$ and $I \vdash_{\mathsf{IL}} E_0$.*

In the presence of the $\rhd$-modality, the following interpolation property suggests itself. Corollary 4.4 follows immediately from Proposition 4.3 and Theorem 1.

Proposition 4.3 *$\vdash_{\mathsf{IL}} D \rhd E$ if and only if $\vdash_{\mathsf{IL}} D \to E \vee \Diamond E$.*

Proof of Proposition 4.3: "$\Leftarrow$" follows from Proposition 2.2.2. For "$\Rightarrow$", assume $\nvdash_{\mathsf{IL}} D \to E \vee \Diamond E$. By completeness there exists an IL-model $\langle\langle W, R, S\rangle, V\rangle$ and some world $w_1 \in W$ such that $w_1 \models D$ and $w_1 \not\models E \vee \Diamond E$. Let $W' \overset{\text{def}}{=} \{w \in W : w_1 R w\} \cup \{w_1, w_0\}$, where w_0 is some fresh element. By R' we denote the transitive closure of $(R\lceil(W' \setminus \{w_0\}) \cup \langle w_0, w_1\rangle)$. Here by $R\lceil(W' \setminus \{w_0\})$ we understand the restriction of the relation R to the set $W' \setminus \{w_0\}$. Let S'_{w_0} be the reflexive closure of $R\lceil(W' \setminus \{w_0\})$, and $S'_w = S_w$, for $w \in W' \setminus \{w_0\}$. The so obtained $\langle\langle W', R', S'\rangle, V'\rangle$, where V' is any valuation extending V, is an IL-model. Moreover, w_1 is an R'-successor of w_0 satisfying D without S'_{w_0}-successor satisfying E. In other words, $w_0 \not\models D \rhd E$. ■

Corollary 4.4 ($\rhd$-Interpolation for IL) *Let D_0, E_0 be IL-formulas. Assume $\vdash_{\mathsf{IL}} D_0 \rhd E_0$. Then there exists an IL-formula $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that $\vdash_{\mathsf{IL}} D_0 \rhd I$ and $\vdash_{\mathsf{IL}} I \rhd E_0$.*

4.2 The Interpolation Theorems for ILP

The system ILP is defined by adding to IL the *persistence principle*, $P : A \rhd B \to \Box(A \rhd B)$ (i.e. if $T + B$ is relatively interpretable in $T + A$, then this can be proved in T). A direct proof of interpolation for ILP can be obtained using the techniques introduced in this paper. More elegantly, the question of interpolation for ILP can be reduced to a corollary of Theorem 1 by observing, as Hájek did in [5], that ILP is *strongly interpretable* in IL.

Definition 4.5 (Strong Interpretation of ILP in IL, [5]) We define the translation $^{\#}$ for a formula A in ILP as follows: for A atomic, $A^{\#}$ is A, $^{\#}$ commutes with Boolean connectives and with $\Box$ and $(B \rhd C)^{\#}$ is $(B^{\#} \rhd C^{\#}) \wedge \Box(B^{\#} \rhd C^{\#})$. ◀

Given the P axiom it is immediate that $\vdash_{\mathsf{ILP}} A \leftrightarrow A^{\#}$.

Proposition 4.6 $\vdash_{\mathsf{IL}} A^{\#}$ *if and only if* $\vdash_{\mathsf{ILP}} A$.

Proof of Proposition 4.6: Left to right is trivial. The other direction is proved by induction on the length of the proof in ILP. The core of the proof consists of establishing that the translation of all the axioms of ILP are theorems of IL. ■

THEOREM 2 (The Arrow Interpolation Theorem for ILP) *Let D_0, E_0 be ILP-formulas. Assume $\vdash_{\mathsf{ILP}} D_0 \to E_0$. Then there exists an ILP- formula $I \in \mathcal{L}_{D_0} \cap \mathcal{L}_{E_0}$ such that $\vdash_{\mathsf{ILP}} D_0 \to I$ and $\vdash_{\mathsf{ILP}} I \to E_0$.*

Proof of Theorem 2: We reduce interpolation for ILP to interpolation for IL. Assume $\vdash_{\mathsf{ILP}} E_0 \to D_0$. Then by Proposition 4.6, $\vdash_{\mathsf{IL}} E_0^{\#} \to D_0^{\#}$. Applying the interpolation result for IL, we then obtain a formula I such that $\vdash_{\mathsf{IL}} E_0^{\#} \to I$ and $\vdash_{\mathsf{IL}} I \to D_0^{\#}$. Obviously then, $\vdash_{\mathsf{ILP}} E_0^{\#} \to I$ and $\vdash_{\mathsf{ILP}} I \to D_0^{\#}$. As $\vdash_{\mathsf{ILP}} A^{\#} \leftrightarrow A$, it follows that $\vdash_{\mathsf{ILP}} E_0 \to I$ and $\vdash_{\mathsf{ILP}} I \to D_0$. Note that I is in the common language of E_0, D_0, since the translation $^{\#}$ does not alter languages. ■

Reasoning as we did for IL it is straightforward to prove that

Corollary 4.7 ILP *has turnstile and $\rhd$-interpolation.*

5 Beth Definability and Fixed Points for (Extensions of) IL

Ever since 1957 when W. Craig gave an alternative proof for the Beth Definability Theorem for first-order logic via interpolation, these two properties are often studied together. Albeit their close relation, it turns out that they behave quite different in the context of interpretability logics.

5.1 From Beth Definability to Fixed Points

Definition 5.1 (Beth Definability Property) A logic $\mathcal{L}$ has the *Beth Definability Property* iff for all formulas $A(\bar{p}, r)$ whose proposition letters occur among $\bar{p}, r$, the following holds:

$$\text{If } \vdash_{\mathcal{L}} \boxdot A(\bar{p}, r) \wedge \boxdot A(\bar{p}, r') \to (r \leftrightarrow r'),$$

(in words, if $A(\bar{p}, r)$ *implicitly defines* r in terms of $\bar{p}$) then there exists a formula $C(\bar{p})$ (called an *explicit definition*) such that

$$\vdash_{\mathcal{L}} \boxdot A(\bar{p}, r) \to (C(\bar{p}) \leftrightarrow r).$$

Here $\boxdot A$ abbreviates $A \wedge \Box A$. ◀

Using a standard argument (cf. e.g. [2]) we can easily derive the Beth definability property for IL from Theorem 1. But as we will shortly see (cf. Theorem 4), we can infer much more.

One of the well-known applications of the Beth definability property can be found in the literature on provability logic. In e.g. [1, 13], C. Smoryński derives for provability logic L the *existence* of fixed points, the more interesting half of the Fixed Point Theorem, from the *uniqueness* of fixed points via an application of the Beth property. Along the same lines we obtain the following result, a direct proof of which was already given by de Jongh and Visser in [10].

THEOREM 3 (Fixed Point Theorem for IL) *Let $A(\bar{p}, r)$ be an IL-formula which is* modalized *in r, i.e., r occurs only in the scope of $\Box$ or $\rhd$. Then there is a formula $F(\bar{p})$ (called a* fixed point*) such that*

$$\vdash_{\mathsf{IL}} \boxdot(r \leftrightarrow A(\bar{p}, r)) \wedge \boxdot(r' \leftrightarrow A(\bar{p}, r')) \rightarrow r \leftrightarrow r' \text{ (uniqueness), and}$$
$$\vdash_{\mathsf{IL}} F(\bar{p}) \leftrightarrow A(\bar{p}, F(\bar{p})) \text{ (existence).}$$

Note that the above theorem obviously implies that *all extensions of* IL have the fixed point property.

5.2 From Fixed Points to Beth Definability

Another angle on the Beth property and fixed points was taken in [11]. There L. Maksimova shows that for provability logics the fixed point property (existence together with uniqueness) in its turn implies the Beth property. In what follows we adapt that proof to interpretability logics.

THEOREM 4 (The Beth Definability Theorem for extensions of IL) *Let $\mathcal{L}$ be an extension of IL. Then $\mathcal{L}$ has the Beth definability property.*

A first difficulty that arises in proving Theorem 4 from Theorem 3, is the more general character of Theorem 4. The Fixed Point Theorem that is at our disposal is a statement about *modalized* formulas, whereas the Beth Theorem is about *arbitrary* formulas. The next lemma reduces arbitrary formulas to ones which are 'largely modalized', and thereby provides a starting point for proving Beth's Theorem from the Fixed Point Theorem.

Lemma 5.2 ([11]) *Let $\mathcal{L}$ be an extension of IL, and let $A(\bar{p}, r)$ be an arbitrary IL-formula. Then there exist IL-formulas $A_1(\bar{p}, r)$, $A_2(\bar{p}, r)$ which are modalized in r such that*

$$\vdash_{\mathcal{L}} A(\bar{p}, r) \leftrightarrow [(r \wedge A_1(\bar{p}, r)) \vee (\neg r \wedge A_2(\bar{p}, r))].$$

This observation was first made by L. Maksimova, and rests on some syntactic considerations: writing an arbitrary formula in disjunctive normal form and collecting the disjuncts containing r and the ones containing $\neg r$ will give the form required by Lemma 5.2. Note that this line of argumentation does not make use of any particularities of interpretability logics as such, and hence the above lemma holds for a much more general class of logics than is stated.

Proof of Theorem 4: Let $\mathcal{L}$ be an arbitrary but fixed extension of IL. Consider an implicit $\mathcal{L}$-definition $A(\bar{p}, r)$ of r in terms of $\bar{p}$. Abbreviating $A(\bar{p}, r)$ to $A(r)$, this can be expressed by

$$\vdash_{\mathcal{L}} \boxdot A(r) \wedge \boxdot A(r') \rightarrow (r \leftrightarrow r'). \tag{11}$$

Let us gather some facts. By the previous lemma, there exist formulas $A_1(r)$, $A_2(r)$ which are modalized in r such that

$$\vdash_{\mathcal{L}} A(r) \leftrightarrow [(r \wedge A_1(r)) \vee (\neg r \wedge A_2(r))]. \tag{12}$$

As $\mathcal{L}$ has the fixed point property (cf. Theorem 3), there exists a formula F_1 build up from proposition letters in $\bar{p}$ which is a fixed point of $A_1(r)$, i.e.,

$$\vdash_{\mathcal{L}} F_1 \leftrightarrow A_1(F_1). \tag{13}$$

Moreover, fixed points are unique. That is,

$$\vdash_{\mathcal{L}} \boxdot(r \leftrightarrow A_1(r)) \rightarrow (r \leftrightarrow F_1). \tag{14}$$

Our aim is to show the following claim.

Claim 5.3 $\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow [\Box(A_1(r) \rightarrow r) \rightarrow (A_1(r) \rightarrow r)]$.

From this claim it can be infered, using Löb's axiom $L4$, that

$$\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow \Box(A_1(r) \rightarrow r). \tag{15}$$

Modus ponens applied to Claim 5.3 and (15) gives that

$$\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow (A_1(r) \rightarrow r). \tag{16}$$

On the other hand, from (12) it is obvious that $\vdash_{\mathcal{L}} A(r) \rightarrow (r \rightarrow A_1(r))$. Hence from (16) we conclude that $\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow (r \leftrightarrow A_1(r))$, and therefore,

$$\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow \boxdot(r \leftrightarrow A_1(r)).$$

From the uniqueness of fixed points (see (14) above), it then follows that

$$\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow (r \leftrightarrow F_1).$$

Ergo, F_1 is an explicit definition of r. What remains is to prove Claim 5.3.

Proof of Claim 5.3: As observed before, $\vdash A(r) \rightarrow (r \rightarrow A_1(r))$, and hence $\vdash \Box A(r) \rightarrow \Box(r \rightarrow A_1(r))$. Therefore,

$$\vdash_{\mathcal{L}} \boxdot A(r) \wedge \Box(A_1(r) \rightarrow r) \rightarrow \Box(r \leftrightarrow A_1(r)). \tag{17}$$

Notational aside: For the course of this proof, the formula $\boxdot A(r) \wedge \Box(A_1(r) \rightarrow r)$ will be denoted by C.

Note that from the uniqueness of fixed points (14) it follows that $\vdash_{\mathcal{L}} \Box(r \leftrightarrow A_1(r)) \rightarrow \Box(r \leftrightarrow F_1)$, hence by (17)

$$\vdash_{\mathcal{L}} C \rightarrow \Box(r \leftrightarrow F_1). \tag{18}$$

In other words, r and F_1 are equivalent under the box-operator (relative to C). In particular

$$\vdash_{\mathcal{L}} C \rightarrow \Box(A(F_1)), \text{ and} \tag{19}$$

$$\vdash_{\mathcal{L}} C \rightarrow (A_1(r) \rightarrow A_1(F_1)), \tag{20}$$

where (20) holds by virtue of A_1 being modalized in r, and (19) by definition of C. Let us note for future reference that from (20) and the fact that F_1 is a fixed point (cf. (13)) for A_1 it follows that

$$\vdash_{\mathcal{L}} C \to (A_1(r) \to F_1), \tag{21}$$

and $\vdash_{\mathcal{L}} C \to [A_1(r) \to (F_1 \wedge A_1(F_1))]$. By (12), this latter implication shows that $\vdash_{\mathcal{L}} C \to (A_1(r) \to A(F_1))$ which together with (19) implies that

$$\vdash_{\mathcal{L}} C \to (A_1(r) \to \boxdot A(F_1)). \tag{22}$$

$A(r)$ being an implicit definition of r (cf. (11)) entails that $\vdash_{\mathcal{L}} \boxdot A(r) \wedge \boxdot A(F_1) \to (r \leftrightarrow F_1)$. From (22) we then derive that

$$\vdash_{\mathcal{L}} C \to (A_1(r) \to (r \leftrightarrow F_1)). \tag{23}$$

By (21), we obtain the claim.
Q.E.D. Claim 5.3.
Q.E.D. Theorem 4.

The above result reveals a huge contrast between interpolation and definability properties for interpretability logics. For example, as was mentioned in the introduction, all systems between ILM_0 and ILM lack interpolation, whereas by Theorem 4 they all have the Beth property.

References

[1] G. Boolos. *The unprovability of consistency.* Cambridge University Press, Cambridge, 1979. An essay in modal logic.

[2] C. Chang and H. Keisler. *Model Theory.* Elsevier Science Publishers B.V., 1977.

[3] W. Craig. Three uses of the Herbrand-Gentzen theorem in relating model theory and proof theory. *Journal of Symbolic Logic*, 22:269–285, 1957.

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

[5] P. Hájek. IL satisfies interpolation. *Unpublished*, 1992.

[6] K. Ignatiev. Failure of interpolation for ILM. *Unpublished.*

[7] K. Ignatiev. Private comunication. *Unpublished*, 1992.

[8] D. de Jongh and G. Japaridze. The logic of provability. In S. Buss, editor, *Handbook of Proof Theory*, pages 475–546. Elsevier Science Publishers B.V., 1998.

[9] D. de Jongh and F. Veltman. Provability logics for relative interpretability. In [12], pages 31–42, 1990.

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

[11] L. Maksimova. Definability theorems in normal extensions of provability logic. *Studia Logica*, 4:495–507, 1989.

[12] P. Petkov, editor, *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria.* Plenum Press, Boston, 1990.

[13] C. Smoryński. Beth's theorem and self–referential statements. In A. Macintyre, L. Pacholski, and J. Paris, editors, *Computation and Proof Theory*, pages 17–36. North-Holland, 1978.

[14] C. Smoryński. *Self-reference and modal logic.* Springer-Verlag, 1985.

[15] A. Visser. Interpretability logic. In [12], pages 175–209, 1990.

[16] A. Visser. An overview of interpretability logic. In M. Kracht, M. de Rijke and H. Wansing, editors, *Advances in modal logic '96*, pages 307–359. CSLI Publications, Stanford, CA, 1997.

[17] D. Zambella. On the proofs of arithmetical completeness of interpretability logic. *The Notre Dame Journal of Formal Logic*, 35:542–551, 1992.