Interpolation, Definability and Fixed Points in Interpretability Logics

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

ABSTRACT. In this article we study interpolation properties 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 result. Furthermore, these properties are extended to the system ILP. Failure of arrow interpolation for ILW is established by providing an explicit counterexample. The related issues of Beth definability and fixed points are also addressed. It will be shown that for a general class of logics the Beth property and the fixed point property are interderivable. This in particular yields alternative proofs for the fixed point theorem for IL (cf. de Jongh and Visser 1991) and the Beth theorem for all provability logics (cf. Maksimova 1989). Moreover, it entails that all extensions of IL have the Beth property.

1 Introduction

Interpretability logics are extensions of provability logics introduced in Visser 1990. In that paper 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 function should satisfy certain further requirements.) The main importance of interpretability logics is that they permit a finer analysis of arithmetical theories than provability logics. For example, whereas 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 (nonstandard) modal logic. In the present article we establish purely theoretical results about systems of interpretability logic, like the interpolation property for IL and ILP. Hereto, a simple modal reading of $\rhd$ over Kripke models suffices.

Acknowledgment We thank Petr Hájek for kindly sharing with us his unpublished partial proof of interpolation for IL (Hájek 1992) and his correspondence with Ignatiev on this matter (Ignatiev 1992). Moreover, we are thankful for his permission to publish his elegant proof that the interpolation property for ILP follows from interpolation of IL (also in Hájek 1992).

1.1 Interpretability and Interpolation

When a new logic is defined, some questions immediately come to mind as a yardstick by which to measure the behavior of the newborn logic. Is the logic sound, complete, decidable? In this article we deal with some of these metalogical questions: Craig interpolation, Beth definability and fixed points. The last two properties —which will be shown to hold for all extensions of the basic system IL— will follow from the interpolation property for IL.

In Craig 1957 the famous interpolation theorem for first-order logic (FO) was proven: 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 (for the definition of the systems mentioned we refer to Visser 1997). In Visser 1997 a proof by Ignatiev of failure of interpolation for ILM is adapted, showing that systems between ILM_0 and ILM do not have interpolation. It follows for example that ILW^* does not have

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

interpolation. In de Rijke 1992 unary interpretability logic, i.e., the logic of $(\top \rhd \psi)$ is studied. de Rijke 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 1992 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 Ignatiev 1992, 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. Hájek 1992). By using the model theoretic notion of bisimulation we will furthermore prove failure of interpolation for ILW.

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	no	no
Proved in	This paper	Hájek 1992 This paper	Ignatiev		This paper	Visser 1997

Unary Systems	il	ilp	ilm			
Interpolation	yes	yes	yes			
Proved in	de Rijke 1992	de Rijke 1992	de Rijke 1992			

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 this topic covering the arithmetical interest of the project we refer to Japaridze and de Jongh 1998 and Visser 1997. 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 $\rhd$-interpolation also hold for IL. We also show that all these properties transfer to ILP. Section 5 provides a counter example to arrow interpolation for ILW. In the final section we will deepen an interesting interplay between Beth definability and fixed points. For a general class of logics these two properties will be shown to be interderivable. This class includes all provability and interpretability logics. Since the Beth property can be derived in IL from arrow interpolation as usual, this yields an alternative proof for the fixed point theorem for IL (cf. de Jongh and Visser 1991). Moreover, it implies that all extensions of the basic system of provability logic L and all extensions of IL have the Beth definability property. This extends the result in Maksimova 1989 concerning the Beth property for provability logics.

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. Visser 1997), a formalization in arithmetic of Henkin's completeness theorem.

In the proof of Theorem 1 the following facts will be useful. The proofs can be found in Japaridze and de Jongh 1998 and Visser 1997.

Proposition 2.2 *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). $\dashv$

We now turn to semantics. A Kripke semantics (in this case also called *Veltman semantics*) for IL was first presented in de Jongh and Veltman 1990.

Definition 2.3 (IL-Frame, IL-Model, 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 $w\uparrow \ =_{\text{def}} \{u \in W : wRu\}$.
  - 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 1990 provides 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 Smoryński 1978. 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 (see de Jongh and Veltman 1990).

Definition 2.5 ($\sim A$, Adequate Set) 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 when appropriate. 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. $\vdash \bigwedge X \to A$ will be written simply as $\vdash X \to A$.

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 central role played by *maximal consistent sets* in the latter is in the former taken over by *complete inseparable 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. Smoryński 1985 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. Consider a world w which contains a formula of the form $D \rhd A$. Hence, by the truth lemma, $w \models D \rhd A$. 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. 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 formalize the intuitive notion of 'A-avoiding successor' would be via the following concept of A-criticality (see Hájek 1992).

Definition 2.12 **(A-Critical, preliminary)** 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. $X_1 =_{\mathrm{def}} \{\neg D, \Box\neg D : D \rhd A \in X\} \subseteq X'$.
   $Y_1 =_{\mathrm{def}} \{\neg E, \Box\neg E : E \rhd A \in Y\} \subseteq Y'$.

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
   $X_1 =_{\mathrm{def}} \{\neg D, \Box\neg D : D \rhd A \in X\} \subseteq X'$.
   $Y_1 =_{\mathrm{def}} \{\neg E, \Box\neg E : \Box\neg E \in \mathcal{A}_Y \,\&$
   $\exists C \in \mathcal{L}_X \cap \mathcal{L}_Y [\vdash Y \to (E \rhd C) \;\&\; \vdash X \to (C \rhd A)]\} \subseteq Y'$.
3. If $\Box\neg A \in \mathcal{A}_Y$, then
   $X_2 =_{\mathrm{def}} \{\neg D, \Box\neg D : \Box\neg D \in \mathcal{A}_X \,\&$
   $\exists C \in \mathcal{L}_X \cap \mathcal{L}_Y [\vdash X \to (D \rhd C) \;\&\; \vdash Y \to (C \rhd A)]\} \subseteq X'$.
   $Y_2 =_{\mathrm{def}} \{\neg E, \Box\neg E : E \rhd A \in Y\} \subseteq Y'$.

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. $\dashv$

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 $\mathcal{M} =_{\text{def}} \langle\langle W, R, S\rangle, V\rangle$ 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:
  - $w_0 =_{\text{def}} [(\langle X_0, Y_0\rangle, [])] \in W$.
  - 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 $X \subseteq \mathcal{A}_{D_0}$, $Y \subseteq \mathcal{A}_{E_0}$ and $\langle X_n, Y_n\rangle \prec_A \langle X, Y\rangle$, 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 $w * [(\langle X', Y'\rangle, \tau_w * [A])] \subseteq u$, $w * [(\langle X'', Y''\rangle, \tau_w * [A])] \subseteq v$.
- 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 $w \in V(p)$ iff $p \in X_w \cup Y_w$.

End of construction.

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 $w'S_ww''$.

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,

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

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 $v \not\models F$. Hence $w \not\models G \rhd F$, 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_Au, for some formula A. By induction hypothesis, $G \in X_u$. By Lemma 3.5 there exists some $v \in W$ such that uS_wv and $F \in X_v$. Again by the induction hypothesis $v \models F$, and it follows that $w \models G \rhd F$. $\dashv$

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

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

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

$$\boxdot X =_{\text{def}} \{D, \Box D : \Box D \in X\}.$$

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 $u =_{\text{def}} w * [(\langle X_u, Y_u\rangle, \tau_w * [F])]$ 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:

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

Henceforth we will simply omit the index set (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 $\boxdot X_w$ 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,

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

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. $\dashv$

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

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

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 $v =_{\text{def}} w * [(\langle X_v, Y_v \rangle, \tau_w * [A])]$ 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

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

Reasoning as before we see that

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

and

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

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

$$(3) \qquad \vdash X_w \to [(F \rhd A) \to (G \rhd A)].$$

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,

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

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

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

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

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

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,

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

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_Au, 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 $v =_{\text{def}} w * [(\langle X_v, Y_v \rangle, \tau_w * [A])]$ has all the required properties. $\dashv$

This ends the proof of Theorem 1. $\dashv$

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 → is replaced by ⊢), and their corresponding semantic versions. There is no general connection between these properties. However, in the presence of 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 $\vdash_{\mathsf{IL}} (A \wedge \Box A) \rightarrow B$.*

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

Corollary 4.2 ($\vdash$-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 $I \vdash_{\mathsf{IL}} E_0$.*

Because the $\rhd$-modality can be thought of as a conditional, 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 $\vdash_{\mathsf{IL}} D \rightarrow E \vee \Diamond E$.*

Proof of Proposition 4.3. "$\Leftarrow$" follows from Proposition 2.2.2. For "$\Rightarrow$", assume $\nvdash_{\mathsf{IL}} D \rightarrow E \vee \Diamond E$. 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 $W' =_{\text{def}} \{w \in W : w_1 R w\} \cup \{w_1, w_0\}$, 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 a 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 : $A \rhd B \rightarrow \Box(A \rhd B)$ (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 was done in Hájek 1992, that ILP is *strongly interpretable* in IL.

Definition 4.5 (Strong Interpretation of ILP in IL) 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* $\vdash_{\mathsf{ILP}} I \to E_0$.

The proof below is due to Hájek.

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 Failure of Interpolation in ILW

We finish the part of this article on interpolation with a negative result: ILW, the system obtained by extending IL with the axiom $W : A \rhd B \to A \rhd (B \wedge \Box\neg A)$, does not have interpolation. To establish this failure we should exhibit a pair of formulas D, E such that $\vdash_{\mathsf{ILW}} D \to E$ whereas no interpolant exists for D and E. We propose the following implication

$$D \to E =_{\text{def}} (\Box(s \leftrightarrow \Box\neg p) \wedge (p \rhd q)) \to (q \rhd r \to r \rhd (r \wedge s)).$$

Claim 5.1 $\vdash_{\mathsf{ILW}} D \to E$.

Proof of Claim 5.1. That $D \to E$ is a theorem of ILW follows from $J2 : p \rhd q \to (q \rhd r \to p \rhd r)$ and (*) $\vdash_{\mathsf{ILW}} p \rhd r \to r \rhd (r \wedge \Box\neg p)$. To prove this last theorem, reason as follows. By propositional logic, $\vdash_{\mathsf{ILW}} r \to ((r \wedge \Box\neg p) \vee (r \wedge \Diamond p))$ and with the aid of $J1$ we derive

$$\vdash_{\mathsf{ILW}} r \rhd ((r \wedge \Box\neg p) \vee (r \wedge \Diamond p)). \tag{11}$$

On the other hand, from $W : p \rhd r \to p \rhd (r \wedge \Box\neg p)$, by $J2$ and $J5$ we obtain $\vdash_{\mathsf{ILW}} p \rhd r \to \Diamond p \rhd (r \wedge \Box\neg p)$ and hence,

$$\vdash_{\mathsf{ILW}} p \rhd r \to (r \wedge \Diamond p) \rhd (r \wedge \Box\neg p). \tag{12}$$

(11) and (12) immediately imply (*). ⊣

What remains to prove is that $D \to E$ does not have an interpolant. The following notion of bisimulation, introduced in Visser 1990, is crucial.

Definition 5.2 (P-bisimulation) Let $\mathcal{M} = \langle\langle W, R, S\rangle, V\rangle$ and $\mathcal{M}' = \langle\langle W', R', S'\rangle, V'\rangle$ be two models and P a set of proposition letters. A *P-bisimulation* between $\mathcal{M}$ and $\mathcal{M}'$ is a nonempty relation $Z \subseteq W \times W'$ such that

atom $wZw' \Rightarrow (w \in V(p)$ iff $w' \in V'(p))$, for all $p \in P$.

zig If wZw' and wRv, then there is a v' with vZv' and $w'R'v'$ and, for all u' with $v'S_{w'}u'$, there is an u with uZu' and vS_wu.

zag If wZw' and $w'R'v'$, then there is an v with vZv' and wRv and, for all u with vS_wu, there is an u' with uZu' and $v'S_{w'}u'$.

Recall that by $\mathcal{L}_P$ we denote the set of IL-formulas built up from proposition letters in P. The important result about P-bisimulations (Visser 1990) is that they preserve truth of formulas in $\mathcal{L}_P$.

Proposition 5.3 *Let $\mathcal{M}$ and $\mathcal{M}'$ be two IL-models and Z a P-bisimulation between them. Then for any formula $A \in \mathcal{L}_P$, $wZw' \Rightarrow (\mathcal{M}, w \Vdash A$ iff $\mathcal{M}', w' \Vdash A)$.*

ILW-frames are IL-frames such that for each w, the composition $R \circ S_w$ is upwards wellfounded. de Jongh and Veltman 1999 proves completeness for ILW w.r.t. finite ILW-models.

Consider now the two ILW-models in Figure 1. We use the following conventions. Worlds are labeled with the proposition letters which hold in them. Filled arrows stand for both R and S relations, while dashed arrows

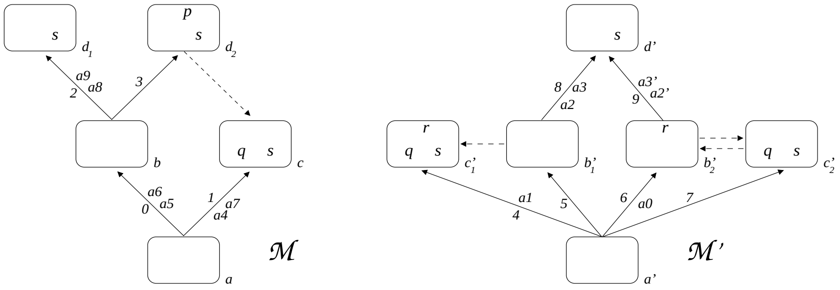

Figure 1: $\{q, s\}$-bisimilar models.

are S relations only. Whenever we have wRv, wRu and vSu then actually vS_wu. Finally we should consider the transitive closure of the filled arrows and the reflexive-transitive closure of the dashed ones.

We claim that the relation linking pairs of worlds labeled by the same letter (disregarding subindices and $'$) is a $\{q, s\}$-bisimulation. Condition **atom** is easily checked. Verifying **zig** and **zag** requires more work.

To this aim, we point out that we can interpret the **zig** (and similarly the **zag**) condition on bisimulations as a rule in a game which requires that whenever wZw' and a 'move' wRv has been played, we should be able to answer with a 'counter-move' $w'R'v'$ which fulfills the necessary conditions on S and Z. We have labeled the arrows in the models according to this idea. An arrow marked n (for $n \in \{0, \ldots, 9\}$) is 'answered' by the arrow marked an in the other model. Note that some arrows are played twice because the bisimulation is not injective. For example, arrow 2 is answered by $a2$ when played from the position bZb'_1 and by $a2'$ when played from bZb'_2.

Once the fact that $\mathcal{M}$ and $\mathcal{M}'$ are $\{q, s\}$-bisimilar has been established, what rests is simple. Suppose $D \to E$ above has an interpolant I. Note that $\mathcal{M}, a \Vdash D$. As $\vdash_{\mathsf{ILW}} D \to I$, we have $\mathcal{M}, a \Vdash I$. But, as shown, there is a $\{q, s\}$-bisimulation linking a and a'. Hence by Proposition 5.3, $\mathcal{M}', a' \Vdash I$. As $\vdash_{\mathsf{ILW}} I \to E$, we have $\mathcal{M}', a' \Vdash E$, which is not the case, proving that no interpolant for $D \to E$ exists.

Finally, let us remark that our failure result is of a less general kind than Visser's result (Visser 1997) mentioned in Section 1. As the model on the right in Figure 1 is not an ILM-model we cannot extend the failure result to all logics between ILW and ILM.

6 Beth Definability and Fixed Points

Ever since 1957 when W. Craig gave an alternative proof for the Beth definability theorem for first-order logic (Beth 1953) via interpolation, these two properties are often studied together. Albeit their close relation, it turns out that they behave quite differently in the context of interpretability logics.

Notation 6.1 In this section we will, if useful, denote formulas in such a way that the proposition letters from which they are built up are displayed. For example, the notation $A(\bar{p}, r)$ implies that the proposition letters that occur in the formula A are among $p_1, \ldots, p_k, r$. Also, for any formula A we will write $\boxdot A$ to abbreviate $A \wedge \Box A$. Moreover, a formula A is said to be *modalized* in r if every occurrence of the proposition letter r in A is in the scope of a modality.

Definition 6.2 (Beth Definability Property) A logic $\mathcal{L}$ has the *Beth definability property* iff for all formulas $A(\bar{p}, r)$ the following holds:

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

(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

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

Using a standard argument (cf. e.g. Chang and Keisler 1990) we can easily derive the Beth definability property for IL from Theorem 2. But as we will shortly see (cf. Corollary 6.8), we can infer much more. Hereto we will make a detour via fixed points.

Definition 6.3 (Fixed Point Property) A logic $\mathcal{L}$ has the *fixed point property* iff for any formula $A(\bar{p}, r)$ which is modalized in r, there exists a formula $F(\bar{p})$ (called a *fixed point*) such that

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

Outline of this section First we will show in Theorem 3 that for a general class of logics the fixed point property can be derived from the Beth property. Second it will be proven in Theorem 4 that for these logics the Beth property is in its turn derivable from the fixed point property. Since any extension of IL is such a logic, we can reason as follows. As noted above, IL has the Beth property. Hence by Theorem 3, IL has the fixed point property. The nature of the fixed point property is such that it is inherited by any extension. Via Theorem 4 we then reach the general conclusion that any extension of IL has the Beth property.

Let us finally note that Theorem 3 and Theorem 4 also apply to all extensions of the provability logic L, hereby subsuming known results in that area (see Smoryński 1978 and Maksimova 1989).

6.1 From Beth Definability to Fixed Points

One of the well-known applications of the Beth definability property can be found in the literature on provability logic. In 1978, C. Smoryński derives for the 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. Theorem 3 generalizes this result.

Theorem 3 *Let $\mathcal{L}$ be a normal modal logic in which*
1. $\vdash_{\mathcal{L}} \Box A \rightarrow \Box\Box A$*,*
2. $\vdash_{\mathcal{L}} \boxdot B \rightarrow (\Box A \rightarrow A)$ *implies* $\vdash_{\mathcal{L}} \boxdot B \rightarrow A$*,*
3. the Beth theorem holds.
Then $\mathcal{L}$ has the fixed point property.

Proof of Theorem 3. Let the logic $\mathcal{L}$ satisfy the conditions in the theorem, and let $A(\bar{p}, r)$ be an $\mathcal{L}$-formula which is modalized in r. For brevity, let us write $A(r)$. As every occurrence of r in A is in the scope of a modality, we have

$$\vdash_{\mathcal{L}} \Box(r \leftrightarrow r') \rightarrow (A(r) \leftrightarrow A(r')).$$

Hence

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

An application of the second condition on the logic $\mathcal{L}$ shows that fixed points of $A(r)$ are unique.

In order to construct a fixed point for this formula, we note that uniqueness of fixed points of $A(r)$ is equivalent to $A(r) \leftrightarrow r$ being an implicit definition of r in terms of $\bar{p}$. As $\mathcal{L}$ has the Beth property, this implies the existence of some formula C built up from propositional variables in $\bar{p}$ such that

$$\vdash_{\mathcal{L}} \boxdot(A(r) \leftrightarrow r) \;\rightarrow\; (r \leftrightarrow C). \tag{13}$$

We will show that C is a fixed point for $A(r)$. We first substitute $A(C)$ for r in (13), yielding

$$(14) \qquad \vdash_{\mathcal{L}} \boxdot(A(A(C)) \leftrightarrow A(C)) \ \rightarrow \ (A(C) \leftrightarrow C).$$

Reasoning in $K4$, we then infer that

$$\vdash_{\mathcal{L}} \Box(A(A(C)) \leftrightarrow A(C)) \ \rightarrow \ \Box(A(C) \leftrightarrow C).$$

That is, $A(C)$ and C are equivalent under the $\Box$-operator, given $\Box(A(A(C)) \leftrightarrow A(C))$. As r is modalized in $A(r)$ this implies that

$$\vdash_{\mathcal{L}} \Box(A(A(C)) \leftrightarrow A(C)) \ \rightarrow \ (A(A(C)) \leftrightarrow A(C)).$$

By the second condition on the logic $\mathcal{L}$ this suffices to conclude that

$$\vdash_{\mathcal{L}} A(A(C)) \leftrightarrow A(C).$$

Hence $\vdash_{\mathcal{L}} \boxdot(A(A(C)) \leftrightarrow A(C))$. Recalling (14) we conclude that

$$\vdash_{\mathcal{L}} A(C) \leftrightarrow C.$$

$\dashv$

Remark 6.4 Consider the following weakening of the second condition in Theorem 3,

$$2'. \quad \vdash_{\mathcal{L}} \Box A \rightarrow A \text{ implies } \vdash_{\mathcal{L}} A.$$

We note that in the above proof the existence of fixed points is actually derived from conditions 1 and 3 together with this weakened version of condition 2. Theorem 3 could therefore be rephrased as saying that any normal modal logic $\mathcal{L}$ has the fixed point property if the following requirements are met: $\vdash_{\mathcal{L}} \Box A \rightarrow \Box\Box A$, condition 2' holds, $\mathcal{L}$ has the Beth property, and fixed points in $\mathcal{L}$ are unique.

Let us verify that Theorem 3 is indeed a generalization of the aforementioned result by Smoryński. Moreover, some more efforts will yield the fixed point theorem for IL, a direct proof of which was already given in de Jongh and Visser 1991.

Corollary 6.5 *Let $\mathcal{L}$ be an extension of L, or an extension of IL. Then $\mathcal{L}$ has the fixed point property.*

Proof of Corollary 6.5. We will check that L and IL satisfy the conditions in Theorem 3. The first condition needs no comment. The Beth theorem for L is proven in Smoryński 1978. As we noted before, the Beth theorem for IL can be derived from Theorem 1 as usual. With regard to the second condition, we note that in any logic $\mathcal{L}$ which satisfies the provability axioms (cf. $L1$–$L4$ in Definition 2.1), $\vdash_{\mathcal{L}} \boxdot B \rightarrow \Box A$ can be inferred from $\vdash_{\mathcal{L}} \boxdot B \rightarrow (\Box A \rightarrow A)$. An application of modus ponens yields condition 2. We conclude from Theorem 3 that L and IL have the fixed point property. This obviously implies that all extensions of L and IL have the fixed point property. $\dashv$

6.2 From Fixed Points to Beth Definability

Another angle on the Beth property and fixed points was first taken in 1989 when L. Maksimova showed that for provability logics the fixed point property in its turn implies the Beth property. In what follows, we will generalize this result.

Theorem 4 *Let $\mathcal{L}$ be a normal modal logic in which*
1. $\vdash_{\mathcal{L}} \Box A \rightarrow \Box\Box A$,
2. $\vdash_{\mathcal{L}} \boxdot B \rightarrow (\Box A \rightarrow A)$ implies $\vdash_{\mathcal{L}} \boxdot B \rightarrow A$,
3. the fixed point theorem holds.
Then $\mathcal{L}$ has the Beth property.

A first difficulty that arises in proving the Beth theorem from the fixed point theorem, is the more general character of the former. For, 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, due to Maksimova (1989), reduces arbitrary formulas to ones which are 'largely modalized', and thereby provides a starting point for proving the Beth theorem from the fixed point theorem.

Lemma 6.6 *Let $\mathcal{L}$ be a normal modal logic, and let $A(\bar{p}, r)$ be an arbitrary $\mathcal{L}$-formula. Then there exist $\mathcal{L}$-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 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 6.6.

Proof of Theorem 4. Let the logic $\mathcal{L}$ satisfy the conditions in the theorem. 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

$$\vdash_{\mathcal{L}} \boxdot A(r) \wedge \boxdot A(r') \rightarrow (r \leftrightarrow r'). \tag{15}$$

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

$$\vdash_{\mathcal{L}} A(r) \leftrightarrow [(r \wedge A_1(r)) \vee (\neg r \wedge A_2(r))]. \tag{16}$$

As $\mathcal{L}$ has the fixed point property, there exists a formula F_1 built up from propositional variables in $\bar{p}$ which is a fixed point of $A_1(r)$, i.e.,

$$\vdash_{\mathcal{L}} F_1 \leftrightarrow A_1(F_1). \tag{17}$$

Moreover, fixed points are unique. Hence,

$$\vdash_{\mathcal{L}} \boxdot(r \leftrightarrow A_1(r)) \rightarrow (r \leftrightarrow F_1). \tag{18}$$

Our aim is to show the following claim.

Claim 6.7 $\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow [\Box(A_1(r) \rightarrow r) \rightarrow (A_1(r) \rightarrow r)]$.

From this claim it follows by the second condition on the logic $\mathcal{L}$ that

$$\vdash_{\mathcal{L}} \boxdot A(r) \rightarrow (A_1(r) \rightarrow r). \tag{19}$$

On the other hand, from (16) it is obvious that $\vdash_{\mathcal{L}} A(r) \rightarrow (r \rightarrow A_1(r))$. Hence from (19) 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 (18) 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 6.7.

Proof of Claim 6.7. As observed before, $\vdash_{\mathcal{L}} A(r) \rightarrow (r \rightarrow A_1(r))$, and hence $\vdash \Box A(r) \rightarrow \Box(r \rightarrow A_1(r))$. Therefore,

$$\vdash_{\mathcal{L}} \boxdot A(r) \wedge \Box(A_1(r) \rightarrow r) \rightarrow \Box(r \leftrightarrow A_1(r)). \tag{20}$$

For notational convenience, let us denote the formula $\boxdot A(r) \wedge \Box(A_1(r) \rightarrow r)$ by C. So (20) amounts to

$$\vdash_{\mathcal{L}} C \to \Box(r \leftrightarrow A_1(r)). \tag{21}$$

From the uniqueness of fixed points (18) it follows that $\vdash_{\mathcal{L}} \Box(r \leftrightarrow A_1(r)) \to \Box(r \leftrightarrow F_1)$, hence by (21)

$$\vdash_{\mathcal{L}} C \to \Box(r \leftrightarrow F_1).$$

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

$$\vdash_{\mathcal{L}} C \to \Box(A(F_1)), \qquad \text{and} \tag{22}$$

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

where (23) holds by virtue of A_1 being modalized in r, and (22) by definition of C. Let us note for future reference that from (23) and the fact that F_1 is a fixed point for A_1 (cf. (17)) it follows that

$$\vdash_{\mathcal{L}} C \to (A_1(r) \to F_1), \tag{24}$$

and $\vdash_{\mathcal{L}} C \to [A_1(r) \to (F_1 \wedge A_1(F_1))]$. By (16), this latter implication shows that $\vdash_{\mathcal{L}} C \to (A_1(r) \to A(F_1))$ which together with (22) implies

$$\vdash_{\mathcal{L}} C \to (A_1(r) \to \boxdot A(F_1)). \tag{25}$$

$A(r)$ being an implicit definition of r (cf. (15)) entails that $\vdash_{\mathcal{L}} \boxdot A(r) \wedge \boxdot A(F_1) \to (r \leftrightarrow F_1)$. From (25) we then derive that

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

By (24), we obtain the claim. $\dashv$

This finishs the proof of Theorem 4. $\dashv$

In the proof of Corollary 6.5 it has already been shown that all extensions of L and all extensions of IL satisfy conditions 1–2 in Theorem 4. Hence from Theorem 4 and Corollary 6.5 we obtain the following result.

Corollary 6.8 *Let $\mathcal{L}$ be an extension of* L*, or an extension of* IL*. Then $\mathcal{L}$ has the Beth property.*

This corollary reveals a striking 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. Or, as was shown in Section 5, ILW does not have this property either. On the other hand, by Corollary 6.8 they all have the Beth property.

References

Beth, E. 1953. On Padoa's method in the theory of definition. *Nederl. Akad. Wetensch. Proc. Ser. A.* **56** = *Indagationes Math.* 15:330–339.

Chang, C., and H. Keisler. 1990. *Model theory.* Studies in Logic and the Foundations of Mathematics, Vol. 73. Amsterdam: North-Holland Publishing Co. Third edition.

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

de Jongh, D., and F. Veltman. 1990. Provability logics for relative interpretability. In *Mathematical Logic.* 31–42. New York: Plenum.

de Jongh, D., and F. Veltman. 1999. Completeness of ILW. Unpublished.

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

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

Hájek, P. 1992. IL satisfies interpolation. Unpublished.

Ignatiev, K. n.d. Failure of interpolation for ILM. Unpublished.

Ignatiev, K. 1992. Private comunication. Unpublished.

Japaridze, G., and D. de Jongh. 1998. The logic of provability. In *Handbook of Proof Theory*, ed. S. Buss. 475–546. Elsevier Science B. V.

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

Petkov, P. (ed.). 1990. *Mathematical logic, Proceedings of the Heyting 1988 summer school in Varna, Bulgaria.* Plenum Press, Boston.

Smoryński, C. 1978. Beth's theorem and self–referential statements. In *Computation and Proof Theory*, ed. A. Macintyre, L. Pacholski, and J. Paris. 17–36. North–Holland, Amsterdam.

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

Visser, A. 1990. Interpretability logic. In *Petkov 1990*, 175–209.

Visser, A. 1997. An overview of interpretability logic. In *Advances in modal logic '96*, ed. Kracht, M., de Rijke, M., and Wansing, H. CSLI Publications, Stanford.