Interpretability Logic

Logic and Applications, IUC, Dubrovnik

Mladen Vuković
vukovic@math.hr
web.math.pmf.unizg.hr/∼vukovic/

Department of Mathematics,
Faculty of Science,
University of Zagreb

September, 2013

The aim of our talk is to give an overview a study of interpretability logic in Zagreb for the last twenty years.

Contents

We introduce our notation and some basic facts, following the article:

A. Visser, *An overview of interpretability logic*, In: Kracht, Marcus (ed.) et al., Advances in modal logic. Vol. 1. Selected papers from the 1st international workshop (AiML'96), Berlin, Germany, October 1996, Stanford, CA: CSLI Publications, CSLI Lect. Notes. 87, pp. 307–359 (1998)

http://www.phil.uu.nl/preprints/lgps/authors/visser/

You can find all details on modal logic in the book:

P. Blackburn, M. de Rijke, Y. Venema, *Modal Logic*, Cambridge University Press, 2001.

A brief overview of history

Provability logic describes the provability predicate of a first-order theory.

Peano arithmetic (or *PA*) is sufficiently strong for coding its language.

We can define Gödel numbers of each symbol, and then express provability predicate *Pr*.

So, the theory *PA* can describe provability predicate and can prove various properties of the predicate *Pr*.

The Gödel sentence G is the fixed point of the predicate $\neg Pr(\cdot)$, i.e. we have $\vdash_{PA} G \leftrightarrow \neg Pr(\lceil G \rceil)$.

L. Henkin asked what we could say about fixed point of the predicate Pr.

Löb's theorem give answer on this question. Here is **Löb's theorem**. For every formula F we have

$$\vdash_{PA} Pr(\lceil F \rceil) \rightarrow F \quad \text{if and only if} \quad \vdash_{PA} F.$$

The natural questions are:

- What can we say about the provability predicate Pr of theory PA?
- Is there a simpler way for investigating it?

Here are the **Bernays–Löb** conditions which express properties of provability predicate which we use in the proof of Gödel's theorem. Here are axioms of some modal system, too.

Bernays–Löb conditions	**Axioms and rules of some modal systems**
$Pr(\lceil A \to B \rceil) \to (Pr(\lceil A \rceil) \to Pr(\lceil B \rceil)$	$\Box(A \to B) \ \to \ (\Box A \ \to \Box B)$
$Pr(\lceil A \rceil) \to Pr(\lceil Pr(\lceil A \rceil) \rceil)$	$\Box A \to \Box \, \Box A$
if A then $Pr(\lceil A \rceil)$	if A then $\Box A$

The idea of treating a provability predicate Pr as a modal operator goes back to Gödel.

The same idea was taken up later by **S. Kripke** and **Montague**, but only in the mid–seventies was the correct choice of axioms, based on Löb's theorem, seriously considered by several logicians independently:

- **G. Boolos**
- **D. de Jongh**
- **R. Magari**
- **G. Sambin**
- **R. Solovay**

Here are the axioms and deduction rules of the system *GL* (**Gödel–Löb system**).

(L0) *all tautologies of the propositional calculus*

(L1) $\Box(A \to B) \to (\Box A \to \Box B)$

(L2) $\Box A \to \Box\,\Box A$

(L3) $\Box(\Box A \to A) \to \Box A$

The deduction rules are modus ponens and necessitation.

An arithmetical interpretation of modal logic is a function $*$ from modal formulas into arithmetical sentences such that:

a) preserving Boolean connectives;

b) $(\Box A)^*$ is $Pr(\ulcorner A^* \urcorner)$.

R. Solovay 1976 proved arithmetical completeness of modal system *GL*. Here is Solovay's first theorem.

Let A be a modal formula. We have:

$$\vdash_{GL} A \quad \text{if and only if} \quad \forall * (\vdash_{PA} A^*).$$

Many theories have equal provability logic – *GL*.

It means that the provability logic *GL* cannot distinguish some properties, as e.g. finite axiomatizability, reflexivity, etc.

After Solovay's results some logicians considered modal representations of other arithmetical properties, for example:
interpretability,
Π_n-conservativity,
interpolability ...

We study interpretability.

Interpretability

Let T be a first-order theory sufficiently strong for coding its sintacs.

Interpretability is a binary relation between extensions of T which are like $T + A$.

We say "$T + A$ is interpretable in $T + B$" if we can "translate" formulas theory $T + A$ such that all theorems of $T + A$ will be theorems theory $T + B$.

Modal logics for interpretability were first studied by **P. Hájek** (1981) and V. **Švejdar** (1983).

A. Visser (1988) introduced the binary modal logic *IL*.

The interpretability logic *IL* results from the provability logic *L*, by adding the binary modal operator $\triangleright$.

For many theories, such as *PA* and its extensions in the same language, the notion of relative interpretability coincides with that of Π_1–conservativity.

The language of the interpretability logic contains:

- propositional letters $p_0,\ p_1, \ldots,$
- logical connectives $\wedge,\ \vee, \rightarrow$ and $\neg,$
- unary modal operator $\Box$
  (we use modal operator $\Diamond$ for abbreviation $\neg\Box\neg$)
- binary modal operator $\triangleright$.

Here are the axioms of the system *IL* (**interpretability logic**).

(L0)–(L3) *axioms of the system GL*

(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) $\Diamond A \rhd A$

The deduction rules of *IL* are modus ponens and necessitation.

The system *IL* is natural from the modal point of view, but arithmetically incomplete.

The system *IL* does not prove any formulas which are valid in every adequate theory.

Various extensions of *IL* are obtained by adding some new axioms.

These new axioms are called the **principles of interpretability**.

Montagna's principle

$$M \equiv A \rhd B \rightarrow (A \wedge \Box C) \rhd (B \wedge \Box C)$$

Denote by *ILM* the system which results from the *IL* by adding the Montagna's principle.

A. Berarducci and **V. Shavrukov** (1990) showed that *ILM* is complete for arithmetical interpretation over *PA*.

The principle of persistency

$$P \equiv A \rhd B \rightarrow \Box(A \rhd B)$$

The system *ILP* results from the *IL* by adding the persistence principle.

A. Visser (1988) obtained the arithmetical completeness for *ILP* over any finitely axiomatizable theory.

So we have at least two different modal logics for interpretability:

one for Peano arithmetic and

one for Gödel–Bernays set theory.

It is still an **open problem** what is the interpretability logic of 'weak' theories like $I\Delta_\circ + EXP$, $I\Delta_\circ + \Omega_1$, PRA...

This is a reason why we have many principles of interpretability.

There are several kinds of semantics for interpretability logic.

The basic semantics are **Veltman models**.

An ordered quadruple $(W, R, \{S_w : w \in W\}, \Vdash)$ is called **Veltman model**, if it satisfies the following conditions:

a) (W, R) is a GL–frame, i.e. W is a non empty set, and the relation R is transitive and reverse well–founded relation on W;

b) For every $w \in W$ is $S_w \subseteq W[w] = \{x : wRx\}$;

c) The relation S_w is reflexive and transitive, for every $w \in W$;

d) If $wRvRu$ then vS_wu;

e) $\Vdash$ is a forcing relation. We emphasize only the definition: $w \Vdash A \rhd B$ if and only if

$$\forall v((wRv \ \& \ v \Vdash A) \Rightarrow \exists u(vS_wu \ \& \ u \Vdash B)).$$

D. de Jongh and F. Veltman proved in 1988 the **modal completeness** of system *IL* with respect Veltman semantics.

Theorem. For each formula F of interpretability logic we have:

$$\vdash_{IL} F \quad \text{if and only if} \quad (\forall W) W \models F$$

§2 Some our results

- Semantics for the Principle M_0
- The Correspondences of Principles in Interpretability Logic
- Bisimulations
- Games
- Normal Forms

Semantics for the principle M_0

We used **generalized Veltman models** for semantics of interpretability principle M_0.

$$M_0 \equiv (A \rhd B) \rightarrow ((\Diamond A \wedge \Box C) \rhd (B \wedge \Box C))$$

We have proved (Glasnik matematički, 1996.) that the following condition determines the principle M_0 on Veltman frames

$$x_1 R x_2 R x_3 \;\&\; x_3 S_{x_1} Y \;\Rightarrow\; \exists Y' \subseteq Y (x_2 S_{x_1} Y' \;\&\; (\forall y \in Y')(\forall z)(yRz \Rightarrow x_2 Rz))$$

The interpretability logic is **not compact** w.r.t. generalized Veltman model.

We proved this by changing Kit, Fine proof for provability logic.

We think that there are two main reasons for other semantics for interpretability logic.

The first reason is complex proofs of arithmetical completeness of interpretability logic.

The second reason is the equal characteristic class of frames for different principles of interpretability.

By using Veltman models (and bisimulations) **V. Švejdar** in 1991. proved independence of principles of interpretability.

We used generalized Veltman models for proving independence of principles of interpretability.

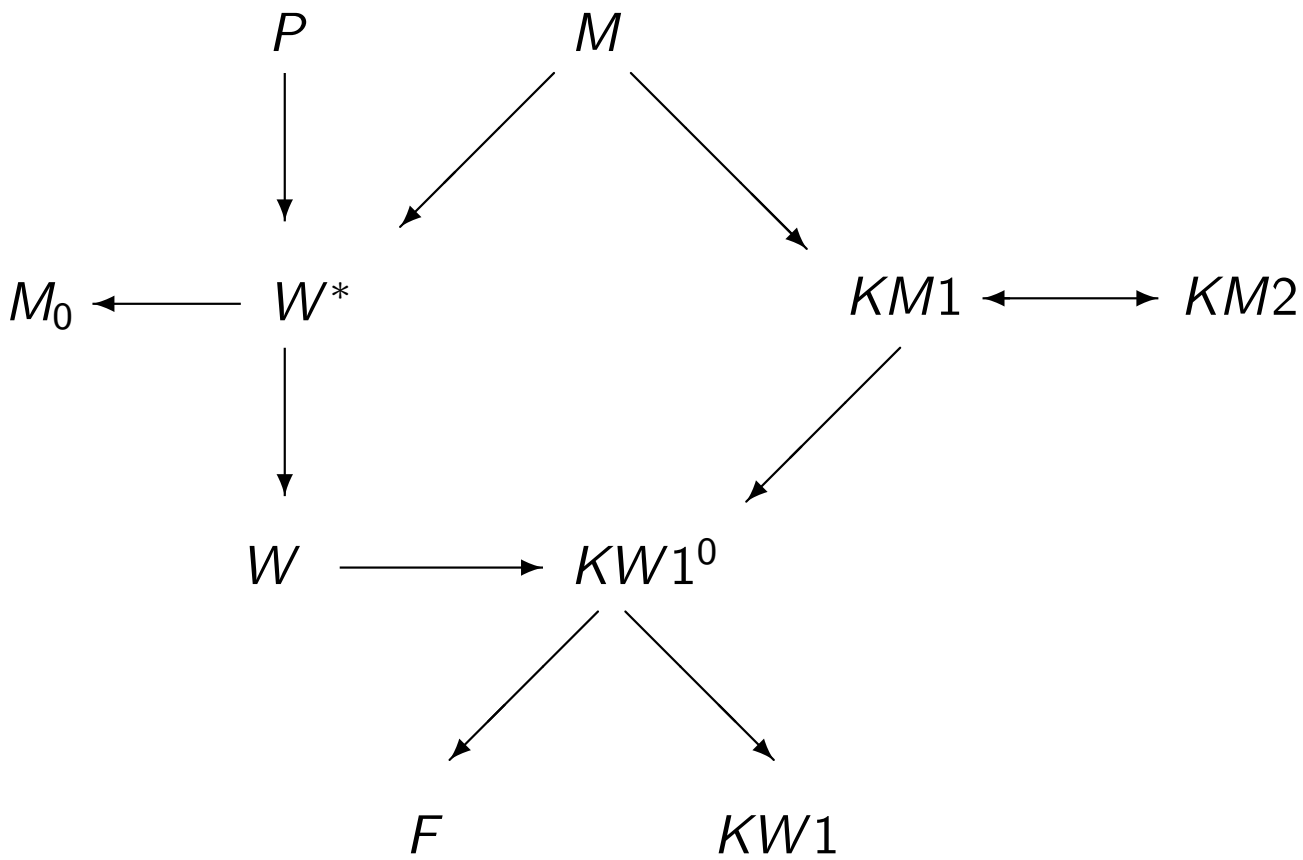
Theorem. (M.V., Notre Dame Jou. For. Log. 1999.)

There are no other implications among combinations of the formulas

$M,\ \ M_0,\ \ KM1,\ \ KM2,\ \ P,\ \ W,\ \ W^*,\ \ KW1^0,\ \ KW1,\ \ F$

except

$M \to W^* \wedge KM1,\ \ P \to W^*,\ \ W^* \to W \wedge M_0,\ \ W \to KW1^0,$
$KM1 \leftrightarrow KM2,\ \ KM1 \to KW1^0$ and $KW1^0 \to F \wedge KW1.$


P
M
M0
W*
KM1
KM2
W
KW1^0
F
KW1

Bisimulations

If we studied correspondence with Kripke models we could consider an isomorphism and an elementary equivalence.

If we want study "weaker" correspondence we could consider bisimulation.

J. van Benthem defined bisimulation of Kripke models.

Roughly speaking, bisimulation is a subset Z of $K \times K_1$, where K and K_1 are some Kripke models.

The basic property of bisimulation is:

$$\text{if } (v, w) \in Z \text{ then } v \Vdash F \text{ iff } w \Vdash F,$$

for all formulas F.

A. Visser (1988) defined bisimulation of Veltman models and proved that every Veltman model with some special property can be bisimulated by a finite Friedman model.

(This fact and de Jongh–Veltman's theorem imply completeness of the system *ILP* w.r.t. finite Friedman models).

A. Berarducci (1990) used a bisimulation for the proof of completeness of system *IL* w.r.t. simplified Veltman models.

By using a bisimulation **A. Visser** proved that Craig interpolation lemma is not true for systems between ILM_0 and ILM.

Definition. A bisimulation between two Veltman models W and W' is a nonempty binary relation $Z \subseteq W \times W'$ such that the following conditions hold:

(at) If wZw' then $W, w \Vdash p$ if and only if $W', w' \Vdash p$, for all propositional variables p;

(forth) If wZw' and wRu, then there exists $u' \in W'$ with $w'R'u'$, uZu' and for all $v' \in W'$ if $u'S_{w'}v'$ there is $v \in W$ such that $uS_w v$;

(back) If wZw' and $w'Ru'$, then there exists $u \in W$ with wRu, uZu' and for all $v \in W$ if $uS_w v$ there is $v' \in W'$ such that $u'S_{w'}v'$.

Bisimulations between (generalized) Veltman models

We defined (Math. Log. Quarterly, 2008) bisimulation between generalized Veltman models, and proved that for a complete image–finite generalized Veltman model W there exists a Veltman model W' that is bisimular to W.

It is an open problem if there is a bisimulation between Veltman model and generalized Veltman model.

Hennessy–Milner theorem

R. de Jonge (2004) proved Hennessy–Milner theorem for Veltman semantics.

We proved (BSL, 2005) Hennessy–Milner theorem for generalized Veltman semantics.

Theorem. Let $\mathbf{W}$ and $\mathbf{W}'$ be two image–finite generalized Veltman models.

Than there is a bisimulation between models $\mathbf{W}$ and $\mathbf{W}'$.

Moreover, if $w \equiv w'$ then $\mathbf{W}, w \underline{\leftrightarrow} \mathbf{W}', w'$.

Bisimulation quotients

Bisimulation quotients and largest bisimulations have been well studied for Kripke models.

We considered how these results extend to Veltman models.

D. Vrgoč and M. V. (Log. Jou. IGPL 2010.; Reports Math. Log. 2011) considered several notions of bisimulation between generalized Veltman models and determined connections between them.

We proved the equivalence between global bisimilarity of models and isomorphisms of their quotients.

Games

We can obtain finite approximations of bisimulations by using games.

V. Čačić and **D. Vrgoč** (Studia Logica, 2013) defined a bisimulation game between Veltman models (two players: defender and spoiler).

Čačić and Vrgoč show that nodes w_1 and w_2 of a Veltman model are bisimilar if and only if defender has a winning strategy in the bisimulation game with the starting configuration (w_1, w_2).

They also define the notion of n–bisimulation between Veltman models and prove the equivalence between the existence of a winning strategy in the n–bisimulation game and the existence of an n–bisimulation.

Normal forms

P. Hájek and V. Švejdar (1991) determined normal forms for the system ILF, and showed that we can eliminate the modal operator $\triangleright$ from IL–formulas.

The normal form for the closed fragment of the interpretability logic IL is an open problem (see Visser, An overview ...)

V. Čačić (Ph.D. 2010 and Math. Comm. 2012) proved that we can eliminate the modal operator $\triangleright$ in some cases.

It is given an example where it is impossible to eliminate $\triangleright$.

V. Čačić, V. Kovač (preprint, 2013) showed that more than 93% of closed IL formulas have a GL–equivalents, and by that, they have the same normal forms as GL formulas.

Correspondence theory

By **van Benthem's characterization theorem**, bisimulation invariance characterizes modal logic as a fragment of first–order logic.

We are interested in corresponding characterization of modal fragments of the first–order language over Veltman models.

Standard translation is a function that maps each modal formula to a first–order formula.

We define a standard translation for interpretability logics.

Let $\sigma = \{P_0, P_1, \ldots\} \cup \{R, S\}$ be a first–order signature, where P_i is an unary relation symbol, R is a binary relation symbol, and S is a ternary relation symbol.

Definition. Let x be a first–order variable. The **standard translation** ST_x taking modal formulas to first–order σ–formulas is defined as follows:

$$ST_x(p_i) = P_i(x)$$

$$ST_x(\neg\varphi) = \neg ST_x(\varphi)$$
$$ST_x(\varphi \wedge \psi) = ST_x(\varphi) \wedge ST_x(\psi)$$

$$ST_x(\Box\varphi) = \forall y(xRy \rightarrow ST_y(\varphi))$$

$$ST_x(\varphi \rhd \psi) = \forall y(xRy \wedge ST_y(\varphi) \rightarrow \exists z(S(x, y, z) \wedge ST_z(\psi)))$$

The following proposition is easy to prove by induction on complexity of modal formula.

Proposition Let φ be a modal formula, W a Veltman model and $w \in W$. Then we have:

$$W, w \Vdash \varphi \qquad \text{if and only if} \qquad W \models ST_x(\varphi)[w].$$

It is easy to prove by induction that bisimilar nodes are modally equivalent, i.e. they satisfy exactly the same *IL*–formulas.

The converse does not hold (see V. Čačić, D. Vrgoč, Studia Logica 2013 for a counterexample), but it does hold for **ω–saturated models**.

In fact, we were able to define an appropriate notion of modal saturation for the language of interpretability logic, which is weaker than ω–saturation, but still sufficient for this converse.

The key step in classical model–theoretic proofs of characterization theorems is construction of a saturated model obtained as an **ultraproduct over a countably incomplete ultrafilter**.

But, an ultraproduct of Veltman models is not necessary a Veltman model itself.

M.V. (Glasnik matematički, 2011) proved that an ultraproduct of Veltman models over *countably complete* ultrafilter is a Veltman model, but this does not imply saturation.

So, the main problem in using classical model theoretic arguments to prove the characterization theorem is to provide the existence of saturated Veltman models.

The cause of these difficulties lies, of course, in the fact that reverse well–foundedness is not a first–order definable property, so it need not be preserved under ultraproducts.

We used bisimulation games (see V. Čačić, D. Vrgoč, Studia Logica 2013) and results from the article A. Dawar, M. Otto, Ann. Pure. App. Log 2009 (**unravelling, locality**).

We proved the following theorem:

Theorem (T. Perkov, M. Vuković, preprint 2013)
A first–order formula is equivalent to standard translation of some formula of interpretability logic with respect to Veltman models if and only if it is invariant under bisimulations between Veltman models.

Filtration

Filtration is a standard method in modal logic.

V. Shetman (Advances in Modal Logic 2004) used a new version of the filtration method which is based on bisimulation. He solved some open problems with finite model property for modal logics.

Can we do this for Veltman models? Can we solve some open problem on finite model property by using filtration?

Fixed points

A. Dawar and M. Otto (Ann. Pure App. Log, 2009.) proved modal characterization theorems over special classes of frames.

As a consequence, they get a generalization of fixed–point theorem for system *GL* (de Jongh–Sambin theorem).

D. de Jongh and F. Veltman (1988) proved fixed–point theorem for interpretability logic.

We will try to get a generalization of the fixed–point theorem for *IL* by using the characterization theorem for Veltman models.