Notre Dame Journal of Formal Logic
Volume 40, Number 2, Spring 1999

The Principles of Interpretability

MLADEN VUKOVIĆ

Abstract A generalized Veltman semantics developed by de Jongh is used to investigate correspondences between several extensions of intepretability logic IL. In this paper we present some new results on independences.

1 Introduction In 1976 Solovay [3] proved arithmetical completeness of modal system L, that is, provability logic. After this some logicians considered modal representations of other arithmetical properties, for example, interpretability, Π_n-conservativity, interpolability, and so on. Modal logics for interpretability were first studied by Hájek [2] and Švejdar [4]. Visser [6] introduced the binary modal logic IL (interpretability logic). The interpretability logic IL results from the provability logic L by adding the binary modal operator $\rhd$.

The arithmetical semantics of interpretability logic is based on the fact that each sufficiently strong theory S contains arithmetical formulas $Pr(x)$ and $Int(x, y)$. Formula $Pr(x)$ expresses that "x is provable in S" (i.e., a formula with Gödel number x is provable in S). Formula $Int(x, y)$ expresses that "$S + x$ interprets $S + y$." An arithmetical interpretation is a function $\star$ from modal formulas into arithmetical sentences preserving Boolean connectives and satisfying

$$(\Box A)^\star = Pr(\ulcorner A^\star \urcorner), \qquad (A \rhd B)^\star = Int(\ulcorner A^\star \urcorner, \ulcorner B^\star \urcorner).$$

($\ulcorner A^\star \urcorner$ denotes Gödel number the formula $A^\star$). A modal formula A is valid in S if $S \vdash A^\star$ for each arithmetical interpretation $\star$. A modal theory T is sound with respect to S if all its theorems are valid in S. A modal theory T is sound if theory T is sound with respect to all reasonable arithmetical theories S. The theory T is complete with respect to S if it proves exactly those formulas that are valid in S. The theory T is complete if it proves exactly those formulas that are valid in any reasonable arithmetical theory S. The soundness of IL was already known and amounts to noticing that all the axioms are valid and the rules of inference preserve validity.

Received May 4, 1998; revised February 17, 2000

Švejdar in [5] investigated independence between principles of interpretability. Švejdar did not consider the principles P, M_0, $KM2$, and W^*. He used Veltman models. Some principles have the same characteristic class of Veltman frames. For example, the principles M and $KM1$ have the same characteristic classes, but characteristic classes of generalzed Veltman frames of these principles are different. The proofs of independence between these principles are relatively complicated by using Veltman semantics.

We proved in [9] that the principles M, P, F, W, W^*, $KM1$, $KM2$, $KW1$, $KW1^\circ$ are not provable in system ILM_0. We used the generalized Veltman semantics as defined by de Jongh. Here we consider all correspondences between the principles.

2 The interpretability logic The language $\mathcal{L}(\Box, \rhd)$ of the interpretability logic contains the propositional letters p_0, $p_1, \dots,$ the logical connectives $\neg, \wedge, \vee, \rightarrow,$ $\longleftrightarrow$, the unary modal operator $\Box$, and the binary modal operator $\rhd$. We use $\bot$ for false and $\top$ for true. The axioms of the interpretability logic IL are:

(L0)	all tautologies of the propositional calculus
(L1)	$\Box(A \rightarrow B) \rightarrow (\Box A \rightarrow \Box B)$
(L2)	$\Box A \rightarrow \Box\Box A$
(L3)	$\Box(\Box A \rightarrow A) \rightarrow \Box A$
(J1)	$\Box(A \rightarrow B) \rightarrow (A \rhd B)$
(J2)	$((A \rhd B) \wedge (B \rhd C)) \rightarrow (A \rhd C)$
(J3)	$((A \rhd C) \wedge (B \rhd C)) \rightarrow ((A \vee B) \rhd C)$
(J4)	$(A \rhd B) \rightarrow (\Diamond A \rightarrow \Diamond B)$
(J5)	$\Diamond A \rhd A$

where $\Diamond$ stands for $\neg\Box\neg$ and $\rhd$ has the same priority as $\rightarrow$. The deduction rules of IL are modus ponens and necessitation.

Axiom (L1) is a formalization of the deduction theorem. Axiom (L2) is an expression of the provable Σ^0_1-completeness of arithmetical theory. Axiom (L3) is a formalization of Löb's theorem. Axioms (J1) – (J3) are clear. Axiom (J4) says that relative interpretability yields relative consistency results. Axiom (J5) is the arithmetized completeness theorem: arithmetical theory plus the assertion that a given theory is consistent interprets the given theory. The system IL is natural from the modal point of view, but arithmetically incomplete. For example, IL does not prove the formula W; that is, $(A \rhd B) \rightarrow (A \rhd (B \wedge \Box(-A)))$, which is 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. From Visser [6] and [7], and Švejdar [5], we have the following principles:

M	$A \rhd B \rightarrow (A \wedge \Box C) \rhd (B \wedge \Box C)$	Montagna's Principle
P	$A \rhd B \rightarrow \Box(A \rhd B)$	Principle of Persistence
M_0	$(A \rhd B) \rightarrow ((\Diamond A \wedge \Box C) \rhd (B \wedge \Box C))$	
F	$(A \rhd \Diamond A) \rightarrow \Box(\neg A)$	Feferman's Principle

W	$(A \rhd B) \to (A \rhd (B \wedge \Box(\neg A)))$	
W^*	$(A \rhd B) \to ((B \wedge \Box C) \rhd (B \wedge \Box C \wedge \Box(\neg A)))$	
$KM1$	$(A \rhd \Diamond B) \to \Box(A \to \Diamond B)$	
$KM2$	$(A \rhd B) \to (\Box(B \to \Diamond C) \to \Box(A \to \Diamond C))$	
$KW1$	$(A \rhd \Diamond \top) \to (\top \rhd (\neg A))$	Transposition Principle
$KW1^\circ$	$((A \wedge B) \rhd \Diamond A) \to (A \rhd (A \wedge (\neg B)))$	

One can naturally pose the question of independence among the quoted principles. Using Veltman models, Švejdar proved the following theorem in [5].

Theorem 2.1 (Švejdar) *No other implications among combinations of the formulas M, $KM1$, W, $KW1^\circ$, $KW1$, F except $M \to W \wedge KM1$, $W \to KW1^\circ$, $KM1 \to KW1^\circ$, and $KW1^\circ \to F \wedge KW1$ are provable over IL.*

In the following theorem we quote Visser's results (see [6] and [7]) about correspondences between the interpretability principles.

Theorem 2.2 (Visser) *We have: $IL(KM1) \vdash KM2$, $ILM \vdash W \wedge KM1$, $ILP \vdash W$, $ILW^* = ILWM_0$, $ILW \nvdash M_0$, $ILM \vdash M_0$, $ILP \vdash M_0$.*

Švejdar in [5] did not investigate the principles P, M_0, $KM2$, and W^*. In this paper we present some new results on independences, that is, we determine all correspondences between the mentioned principles. De Jongh defined generalized Veltman models. Using generalized Veltman models we can show our main result. More precisely, the aim of this paper is to prove the following theorem.

Theorem 2.3 *There is no other implication among combinations of the formulas M, M_0, $KM1$, $KM2$, P, W, W^*, $KW1^\circ$, $KW1$, F except $M \to W^* \wedge KM1$, $P \to W^*$, $W^* \to W \wedge M_0$, $W \to KW1^\circ$, $KM1 \longleftrightarrow KM2$, $KM1 \to KW1^\circ$, and $KW1^\circ \to F \wedge KW1$.*

By picturing we get

$$
\begin{tikzcd}
 & P \arrow[d] & & M \arrow[dll] \arrow[drr] & & & \\
M_0 & W^* \arrow[l] \arrow[d] & & & & KM1 \arrow[r, leftrightarrow] \arrow[dll] & KM2 \\
 & W \arrow[rr] & & KW1^\circ \arrow[dl] \arrow[dr] & & & \\
 & & F & & KW1 & &
\end{tikzcd}
$$

Our theorem follows in a series of propositions and corollaries in Sections 5 and 6.

3 The generalized Veltman semantics Now we define the generalized Veltman semantics for the interpretability logic.

Definition 3.1 (de Jongh) An ordered triple $(W, R, \{S_w : w \in W\})$ is called an IL_{set}-frame and denoted by $\boldsymbol{W}$ if we have

1. (W, R) is an L-frame; that is, W is a nonempty set, and R is a transitive and reverse well-founded relation on W;
2. Every $w \in W$ satisfies
$$S_w \subseteq W[w] \times \mathcal{P}(W[w]) \backslash \{\varnothing\},$$
where $W[w]$ denotes the set $\{x : wRx\}$;
3. The relation S_w is quasi-reflexive for every $w \in W$; that is, wRx implies $xS_w\{x\}$;
4. The relation S_w is quasi-transitive for every $w \in W$; that is, if xS_wY and $(\forall y \in Y)(yS_wZ_y)$ then $xS_w(\cup_{y\in Y} Z_y)$;
5. If $wRuRv$, then $uS_w\{v\}$;
6. If xS_wY and $Y \subseteq Z \subseteq W[w]$, then xS_wZ.

When presenting an IL_{set}-frame by picture, solid arrows indicate R while dotted ones indicate S_w. The relations between nodes (transitivity of the relation R; $wRvRu \Longrightarrow vS_w\{u\}$; quasi reflexivity and quasi transitivity of S_w; condition (6) in the definition of IL_{set}-frame) will not be indicated by arrows.

Definition 3.2 (de Jongh) An ordered quadruple $(W, R, \{S_w : w \in W\}, \Vdash)$ is called the IL_{set}-model (generalized Veltman model) and denoted by $\boldsymbol{W}$ if we have

1. $(W, R, \{S_w : w \in W\})$ is an IL_{set}-frame;
2. $\Vdash$ is the forcing relation between elements of W and formulas of IL, which satisfies the following:
   (a) $w \Vdash \top$ and $w \nVdash \bot$ are valid for every $w \in W$;
   (b) $\Vdash$ commutes with the Boolean connectives;
   (c) $w \Vdash \Box A$ if and only if $\forall x(wRx \Longrightarrow x \Vdash A)$;
   (d) $w \Vdash A \rhd B$ if and only if
$$\forall v((wRv \mathbin{\&} v \Vdash A) \implies \exists V(vS_wV \mathbin{\&} (\forall x \in V)(x \Vdash B))).$$

As usual we shall use the same letter $\boldsymbol{W}$ for a model and a frame. If $\boldsymbol{W}$ is an IL_{set}-frame and A is a formula of IL, we write $\boldsymbol{W} \models A$ if and only if $w \Vdash A$ for all forcing relations $\Vdash$ on $\boldsymbol{W}$ and all nodes w of W. For a modal scheme A and an IL_{set}-frame $\boldsymbol{W}$, $\boldsymbol{W} \models A$ denotes the fact that $\boldsymbol{W} \models B$ for an arbitrary instance B of A. Analogously, we define $\boldsymbol{W} \models A$, if $\boldsymbol{W}$ is an IL_{set}-model . If $\boldsymbol{W}$ is an IL_{set}-model, $V \subseteq W$ and A a formula, the notation $V \Vdash A$ means that $v \Vdash A$ for any $v \in V$.

It is easy to check the adequacy of the system IL with respect to IL_{set}-models. In [9] we proved the completeness of the system IL with respect to generalized Veltman models. We will not define here regular Veltman models (for examples, see [8]).

4 The characteristic classes Let Γ be a set of modal formulas. We will say that an IL_{set}-frame $\boldsymbol{W} = (W, R, \{S_w : w \in W\})$ is in the characteristic class of generalized Veltman frames of Γ if we have $W \models \Gamma$. By $Char_{set}(\Gamma)$ we denote the characteristic class of Γ. Analogously, we denote by $Char(\Gamma)$ the characteristic class of regular Veltman frames of the set Γ. The characteristic class of a principle of interpretability is the characteristic class of the set of all instances of the principle.

Verbrugge determined in an unpublished paper the characteristic classes of the principles P, M, and $KM1$. Denote by (P) the following property of an IL_{set}-frame:

$$x_3 S_{x_1} Y \;\&\; x_1 R x_2 R x_3 \Longrightarrow (\exists Y' \subseteq Y)(x_3 S_{x_2} Y').$$

Then we have $Char_{set}(P) = \{\boldsymbol{W} : IL_{set}\text{-frame } \boldsymbol{W} \text{ possesses the property } (P)\}$. By $(KM1)$ we denote the condition

$$x_2 S_{x_1} Y \Longrightarrow (\exists y \in Y)(\forall z)(yRz \Longrightarrow x_2 Rz).$$

Then we have $Char_{set}(KM1) = \{\boldsymbol{W} : IL_{set}\text{-frame } \boldsymbol{W} \text{ possesses the property } (KM1)\}$. By (M) we denote the following condition:

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

Then we have $Char_{set}(M) = \{\boldsymbol{W} : IL_{set}\text{-frame } \boldsymbol{W} \text{ possesses the property } (M)\}$.

Let (M_0) be the following condition of a generalized Veltman frame:

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

In [9] we proved that $Char_{set}(M_0) = \{\boldsymbol{W} : IL_{set}\text{-frame } \boldsymbol{W} \text{ possesses the property } (M_0)\}$.

It is easy to see that $ILW \vdash F$. Švejdar proved $ILF \nvdash W$. But Švejdar proved in [5] that $Char(F) = Char(W)$. So regular Veltman frames do not distinguish principles F and W. We determined in [10] the characteristic class of generalized Veltman frames of principle F. First, we define some special relations. Let $(W, R, \{S_w : w \in W\})$ be an IL_{set}-frame and let w be its element. With $\overline{S_w}$ and $\overline{R_w}$ we denote the following relations:

1. for $\emptyset \neq A \subseteq W[w]$ and $\mathcal{B} \subseteq \mathcal{P}(W[w]) \backslash \{\emptyset\}$ is valid

$$A \overline{S_w} \mathcal{B} \quad \text{iff} \quad (\forall a \in A)(\exists B \in \mathcal{B})(a S_w B);$$

2. for $C \subseteq \mathcal{P}(W[w]) \backslash \{\emptyset\}$ and $\emptyset \neq D \subseteq W[w]$ is valid

$$C \overline{R_w} D \quad \text{iff} \quad (\forall C \in C)(\forall c \in C)(\exists d \in D)(cRd).$$

We have $Char_{set}(F) = \{\boldsymbol{W}$: relation $\overline{S_w} \circ \overline{R_w}$ is reverse well-founded for all $w \in W\}$. In [11] we proved that $Char_{set}(F) \neq Char_{set}(W)$. We have already mentioned that $Char(M) = Char(KM1)$ and $Char_{set}(M) \neq Char_{set}(KM1)$. So we think the generalized Veltman semantics better distinguishes the principles of interpretability with respect to the characteristic classes.

5 ***The theories* ILP, ILM$_\mathbf{0}$, ILW**, and* IL(KM2)** In a series of propositions and corollaries we determine the correspondences of the theories ILP, ILM_0, ILW^*, and $IL(KM2)$ by the principles M, $KM1$, W, $KW1^\circ$, $KW1$, and F.

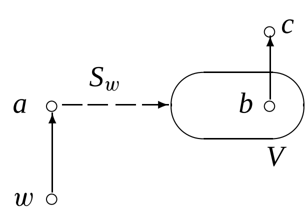
Proposition 5.1 *We have $IL(KM2) \vdash KM1$.*

Proof: Let A and B be arbitrary formulas of the language $\mathcal{L}(\Box, \rhd)$. We write an instance of scheme $KM2$. We substitute the letter B in the scheme $KM2$ by the formula $\Diamond B$ and the letter C by the formula B. Then we have $IL(KM2) \vdash (A \rhd \Diamond B) \to (\Box(\Diamond B \to \Diamond B) \to \Box(A \to \Diamond B))$. This implies $IL(KM2) \vdash (A \rhd \Diamond B) \to \Box(A \to \Diamond B)$. $\Box$

Proposition 5.1 and Visser's theorem imply that the principles $KM1$ and $KM2$ are equivalent over IL. In the rest of the paper we do not investigate the principle $KM2$. Especially, when we deal with principles of the second group we mean on the principles P, W^*, and M_0.

Proposition 5.2 *We have $ILP \nvdash KM1$.*

Proof: We define the IL_{set}-frame $\boldsymbol{W}$ which satisfies the condition (P) and at the same time it does not possess the property $(KM1)$. Let



First we prove that $\boldsymbol{W}$ satisfies the condition (P). Because (P) contains the condition $x_1 R x_2 R x_3$ we consider only the case when $wRbRc$ and $cS_w Y$, where Y is a nonempty subset of $W[w]$. Then the set Y contains the node c. So we have $cS_b Y'$ for the set $Y' = \{c\}$.

It remains to prove that the IL_{set}-frame $\boldsymbol{W}$ does not satisfy the conditon $(KM1)$. We have $aS_w V$ and bRc, but aRc is false. So there is not $y \in V$ such that yRz implies aRz, for all $z \in W$. $\Box$

Visser's theorem and Proposition 5.2 imply $ILP \nvdash M$. Also by Visser's theorem we have $ILP \vdash W^*$. Švejdar's theorem implies $ILP \vdash KW1^\circ \wedge KW1 \wedge F$. In [9] we proved the following theorem. By this theorem, the correspondences of the system ILM_0 with all other principles is completely described.

Theorem 5.3 *The principles M, P, F, W, W^*, $KM1$, $KM2$, $KW1$, $KW1^\circ$ are not provable in ILM_0.*

Many correspondences between the system ILW^* and principles of interpretability follow by means of Visser's result $ILW^* = ILWM_0$ and Švejdar's theorem; that is,

$$ILW^* \vdash W \wedge M_0 \wedge KW1^\circ \wedge KW1 \wedge F.$$

In the following propositions and the corollary we prove the independence between the system ILW^* and the principles $KM1$, M, and P. We use regular Veltman semantics in proofs.

Proposition 5.4 *The principle $KM1$ is not provable in the system ILW^*.*

Proof: We have $Char(KM1) = \{\mathbf{W} : \forall x(S_x \circ R \subseteq R)\}$ and $Char(W^*) = \{\mathbf{W} : \forall x(R \circ S_x \circ R \subseteq R)\}$ (see [8]). Let $W = \{w, a, b, c\}$, $W[w] = \{a, b, c\}$, $W[b] = \{c\}$, and $W[a] = W[c] = \varnothing$. We define the relation S_w by aS_wb. It is easy to check that we have $(W, R, S) \in Char(W^*) \backslash Char(KM1)$. $\square$

Corollary 5.5 *We have $ILW^* \nvdash M$.*

Proof: Visser's theorem and Proposition 5.4 imply the assertion of the corollary. $\square$

Proposition 5.6 *The principle P is not provable in the system ILW^*.*

Proof: We have $Char(P) = \{\mathbf{W} : \forall w \forall x \forall y \forall z(xS_wy$ and $wRzRx$ imply $xS_zy)\}$ (see [8]). Let $W = \{w, a, b, c\}$, $W[w] = \{a, b, c\}$, $W[a] = \{b\}$, and $W[b] = W[c] = \varnothing$. We define the relation S_w by bS_wc. It is easy to check that we have $(W, R, S) \in Char(W^*) \backslash Char(P)$. $\square$

6 *The theories* ILM, IL(KM1), ILW, IL(KW1°), IL(KW1), *and* ILF De Jongh and Veltman in [1] proved $ILM \nvdash P$. Visser's theorem implies $ILM \vdash W^*$.

Corollary 6.1 *We have $ILW \nvdash P$, $ILW \nvdash M_0$, and $ILW \nvdash W^*$.*

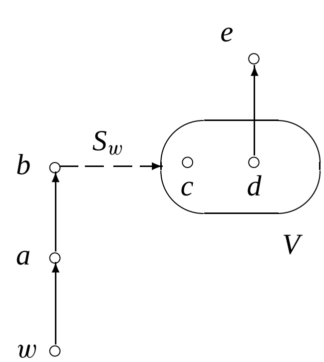
Proof: By Proposition 5.6 and Visser's theorem we get $ILW \nvdash P$. The remaining claims follow from Visser's theorem. $\square$

Corollary 6.2 *Let ILS denote the system $IL + S$, where S is some of principle $KW1^\circ$, $KW1$, and F. Then we have $ILS \nvdash P$, $ILS \nvdash M_0$, and $ILS \nvdash W^*$.*

Proof: In Švejdar's theorem we have $ILW \vdash KW1^\circ \wedge KW1 \wedge F$. Hence, using Corollary 6.1 the assertion of corollary follows. $\square$

Proposition 6.3 *The principle M_0 is not provable in the system $IL(KM1)$.*

Proof: We define the IL_{set}-frame which possesses the property $(KM1)$ and does not possess the property (M_0). As usual we define this IL_{set}-frame by a picture:



First we prove that the defined IL_{set}-frame does not possess the property (M_0). We have $wRaRb$, bS_wV, and aS_wV, but there is no proper subset Y' of the set V such that aS_wY'. Also aS_wV and dRe, but aRe is false.

Now we prove that our IL_{set}-frame possesses the property $(KM1)$. For all $x \in W\backslash\{w\}$ the set $W[x]$ is empty or has only one element. Also we consider only the case $x_1 = w$.

Now we consider all the cases with respect to node x_2. Let $x_2 = a$. If the set Y contains the node a then the condition $(KM1)$ is true. If the set Y contains some R-terminal node (b, c, or e) the condition $(KM1)$ is true again. We emphasize that $aS_w\{d\}$ is not true. Now let $x_2 = b$. Then the set Y contains the node b or it contains c, but the nodes b and c are R-terminal. If $x_2 = c$ then it has to be $c \in Y$. If $x_2 = d$ then the set Y contains the node d or it contains e. If $x_2 = e$ then it has to be $d \in Y$. So in all the cases the condition $(KM1)$ is true. □

Corollary 6.4 *We have $IL(KM1) \nvdash P$ and $IL(KM1) \nvdash W^*$.*

Proof: Visser's theorem and Proposition 6.3 imply $IL(KM1) \nvdash P$. Using Visser's theorem and Proposition 6.3 again, we have $IL(KM1) \nvdash W^*$. □

Acknowledgments The author would like to express his gratitude to Professor A. Visser for valuable comments.

REFERENCES

[1] de Jongh, D., and F. Veltman, "Provability logics for relative interpretability," pp. 31–42 in *Mathematical Logic*, Proceedings of the 1988 Heyting Conference, edited by P. P. Petkov, Plenum Press, New York, 1990. Zbl 0794.03026 MR 92d:03011 6

[2] Hájek, P., "Interpretability in theories containing arithmetic II," *Commentationes Mathematicae Universitatis Carolinae*, vol. 22 (1981), pp. 667–88. Zbl 0487.03032 MR 83j:03094 1

[3] Solovay, R. M., "Provability interpretations of modal logic," *Israel Journal of Mathematics*, vol. 25 (1976), pp. 287–304. Zbl 0352.02019 MR 56:15369 1

[4] Švejdar, V., "Modal analysis of generalized Rosser sentences," *The Journal of Symbolic Logic*, vol. 48 (1983), pp. 986–99. Zbl 0543.03010 MR 85k:03041 1

[5] Švejdar, V., "Some independence results in interpretability logic," *Studia Logica*, vol. 50 (1991), pp. 29–38. Zbl 0728.03016 MR 93c:03026 1, 2, 2, 2, 4

[6] Visser, A., "Interpretability logic," pp. 175–210 in *Mathematical Logic*, Proceedings of the 1988 Heyting Conference, edited by P. P. Petkov, Plenum Press, New York, 1990. Zbl 0793.03064 MR 93k:03022 1, 2, 2

[7] Visser, A., "The formalization of interpretability," *Studia Logica*, vol. 50 (1991), pp. 81–105. Zbl 0744.03023 MR 93f:03009 2, 2

[8] Visser, A., "An overview of interpretability," pp. 307–59 in *Advances in Modal Logic*, edited by M. Kracht, M. de Rijke, and H. Wansing, CSLI Publications, Stanford, 1996. Zbl 0915.03020 MR 1688529 3, 5, 5

[9] Vuković, M., "Some correspondences of principles in interpretability logic," *Glasnik Matematički*, vol. 31 (1996), pp. 193–200. Zbl 0871.03043 MR 98i:03076 1, 3, 4, 5

[10] Vuković, M., "Interpretability logic and generalized Veltman models," abstract, *The Bulletin of Symbolic Logic*, vol. 6 (2000), p. 131. 4

[11] Vuković, M., "Characteristic classes and bisimulations of generalized Veltman models," *Grazer Mathematische Berichte*, vol. 341 (1999), pp. 7–16. Zbl 01606606 MR 1816205 4

Department of Mathematics
University of Zagreb
Bijenička cesta 30
10000 Zagreb
CROATIA
email: vukovic@math.hr