Interpretability in PRA

Marta Bilkova[†], Dick de Jongh[*], and Joost J. Joosten[*],

[*]Institute for Logic Language and Computation
University of Amsterdam
and
[†]Department of Logic
Charles University; Prague

14th July 2007

- We all use the notion $T \rhd S$: T interprets S

- We all use the notion $T \rhd S$: T interprets S
- $T \rhd S$ means (modulo some technical details)

- We all use the notion $T \rhd S$: T interprets S
- $T \rhd S$ means (modulo some technical details)
- $\exists j\ \forall \varphi (\mathrm{Axiom}_S(\varphi) \rightarrow \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$

- We are interested in the structural behavior of the notion of interpretability.

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for
- $T + \varphi \rhd T + \psi$

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for
- $T + \varphi \rhd T + \psi$
- We are interested in the interpretability logic of a theory T:

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for
- $T + \varphi \rhd T + \psi$
- We are interested in the interpretability logic of a theory T:
- The set of all model propositional logical formulas in the language $\Box, \rhd$ which are true regardless how you interpret the variables as arithmetical sentences

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for
- $T + \varphi \rhd T + \psi$
- We are interested in the interpretability logic of a theory T:
- The set of all model propositional logical formulas in the language $\Box, \rhd$ which are true regardless how you interpret the variables as arithmetical sentences
- Of course, reading $\rhd$ as $\rhd_T$, etc.

- We are interested in the structural behavior of the notion of interpretability.
- Interpretability can easily be formalized/arithmetized.
- We shall consider sentential extensions of a base theory
- $\varphi \rhd_T \psi$ stands for
- $T + \varphi \rhd T + \psi$
- We are interested in the interpretability logic of a theory T:
- The set of all model propositional logical formulas in the language $\Box, \rhd$ which are true regardless how you interpret the variables as arithmetical sentences
- Of course, reading $\rhd$ as $\rhd_T$, etc.
- Example: $(\varphi \rhd \psi) \wedge (\psi \rhd \chi) \rightarrow (\varphi \rhd \chi)$

- For all theories T, $\mathsf{IL}(T)$ contains some sort of core logic IL

- For all theories T, $\mathrm{IL}(T)$ contains some sort of core logic IL
- $\mathrm{IL}(T)$ is characterized by some additional axiom schemes on top of that

- For all theories T, $\mathrm{IL}(T)$ contains some sort of core logic IL
- $\mathrm{IL}(T)$ is characterized by some additional axiom schemes on top of that
- For example, for theories with full induction, we have that *Montagna's Axiom* holds

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

- For all theories T, $\mathrm{IL}(T)$ contains some sort of core logic IL
- $\mathrm{IL}(T)$ is characterized by some additional axiom schemes on top of that
- For example, for theories with full induction, we have that *Montagna's Axiom* holds

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

- It turns out that precisely ILM is, e.g. IL(PA) (Shavrukov 1988; Berarducci 1990)

- For all theories T, $\mathrm{IL}(T)$ contains some sort of core logic IL
- $\mathrm{IL}(T)$ is characterized by some additional axiom schemes on top of that
- For example, for theories with full induction, we have that *Montagna's Axiom* holds

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

- It turns out that precisely ILM is, e.g. IL(PA) (Shavrukov 1988; Berarducci 1990)
- Likewise, the interpretability logic for finitely axiomatized theories is known

- For all theories T, $\mathrm{IL}(T)$ contains some sort of core logic IL
- $\mathrm{IL}(T)$ is characterized by some additional axiom schemes on top of that
- For example, for theories with full induction, we have that *Montagna's Axiom* holds

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

- It turns out that precisely ILM is, e.g. IL(PA) (Shavrukov 1988; Berarducci 1990)
- Likewise, the interpretability logic for finitely axiomatized theories is known
- And no other!

- For all theories T, IL(T) contains some sort of core logic IL
- IL(T) is characterized by some additional axiom schemes on top of that
- For example, for theories with full induction, we have that *Montagna's Axiom* holds

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

- It turns out that precisely ILM is, e.g. IL(PA) (Shavrukov 1988; Berarducci 1990)
- Likewise, the interpretability logic for finitely axiomatized theories is known
- And no other!
- That's were PRA comes in

- Consider again

$$\exists j\ \forall \varphi(\mathrm{Axiom}_S(\varphi) \rightarrow \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Consider again

$$\exists j\ \forall \varphi(\mathrm{Axiom}_S(\varphi) \to \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Certainly Σ_3

- Consider again

$$\exists j\ \forall \varphi(\mathrm{Axiom}_S(\varphi) \rightarrow \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Certainly Σ_3
- When S has finitely many axioms, then Σ_1

- Consider again

$$\exists j \ \forall \varphi(\mathrm{Axiom}_S(\varphi) \rightarrow \exists p \ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Certainly Σ_3
- When S has finitely many axioms, then Σ_1
- When T is reflexive, then Π_2. (Orey-Hájek).

- Consider again

$$\exists j\ \forall \varphi(\mathrm{Axiom}_S(\varphi) \to \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Certainly Σ_3
- When S has finitely many axioms, then Σ_1
- When T is reflexive, then Π_2. (Orey-Hájek).
- When T is reflexive, we have access to Montagna's Principle:

$$(T \rhd S) \to ((T \wedge \Box\gamma) \rhd (S \wedge \Box\gamma))$$

- Consider again

$$\exists j\ \forall \varphi(\mathrm{Axiom}_S(\varphi) \rightarrow \exists p\ \mathrm{Proof}_T(p, \ulcorner \varphi^j \urcorner))$$

- Certainly Σ_3
- When S has finitely many axioms, then Σ_1
- When T is reflexive, then Π_2. (Orey-Hájek).
- When T is reflexive, we have access to Montagna's Principle:

$$(T \rhd S) \rightarrow ((T \wedge \Box\gamma) \rhd (S \wedge \Box\gamma))$$

- Every extension of PRA by Σ_2 sentences is reflexive (Parsons, Beklemishev, etc)

- Consider again

$$\exists j\ \forall\varphi(\mathrm{Axiom}_S(\varphi) \to \exists p\ \mathrm{Proof}_T(p, \ulcorner\varphi^j\urcorner))$$

- Certainly Σ_3
- When S has finitely many axioms, then Σ_1
- When T is reflexive, then Π_2. (Orey-Hájek).
- When T is reflexive, we have access to Montagna's Principle:

$$(T \rhd S) \to ((T \wedge \Box\gamma) \rhd (S \wedge \Box\gamma))$$

- Every extension of PRA by Σ_2 sentences is reflexive (Parsons, Beklemishev, etc)
- $(\alpha \rhd_{\mathrm{PRA}} \beta) \to ((\alpha \wedge \Box\gamma) \rhd_{\mathrm{PRA}} (\beta \wedge \Box\gamma))$
  whenever $\alpha \in \Sigma_2$

- $\mathsf{B} := (A \rhd B) \to (A \wedge \Box C) \rhd (B \wedge \Box C) \qquad \text{for } A \in \mathsf{ES}_2$

- $\mathsf{B} := (A \rhd B) \rightarrow (A \wedge \Box C) \rhd (B \wedge \Box C) \qquad \text{for } A \in \mathsf{ES}_2$
- where

- $\mathsf{B} := (A \rhd B) \to (A \wedge \Box C) \rhd (B \wedge \Box C) \quad$ for $A \in \mathsf{ES}_2$
- where
- $$\mathsf{ES}_2 \quad := \quad \Box\mathcal{A} \mid \neg\Box\mathcal{A} \mid \mathsf{ES}_2 \wedge \mathsf{ES}_2 \mid \mathsf{ES}_2 \vee \mathsf{ES}_2 \mid \neg(\mathsf{ES}_2 \rhd \mathcal{A})$$

- If T and S are Π_2 axiomatized theories with

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$
- then, $T \equiv_1 (T \cup S)$

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$
- then, $T \equiv_1 (T \cup S)$
- So,
$$(\alpha \rhd \beta) \wedge (\beta \rhd \alpha) \rightarrow (\alpha \rhd (\alpha \wedge \beta))$$
whenever,

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$
- then, $T \equiv_1 (T \cup S)$
- So,

$$(\alpha \rhd \beta) \wedge (\beta \rhd \alpha) \rightarrow (\alpha \rhd (\alpha \wedge \beta))$$

  whenever,
- $\alpha,\ \beta \in \Sigma_2$, and

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$
- then, $T \equiv_1 (T \cup S)$
- So,

$$(\alpha \rhd \beta) \wedge (\beta \rhd \alpha) \to (\alpha \rhd (\alpha \wedge \beta))$$

  whenever,
- $\alpha,\ \beta \in \Sigma_2$, and
- $\alpha,\ \beta \in \Pi_2$.

- If T and S are Π_2 axiomatized theories with
- $T \equiv_1 S$
- then, $T \equiv_1 (T \cup S)$
- So,
$$(\alpha \rhd \beta) \wedge (\beta \rhd \alpha) \rightarrow (\alpha \rhd (\alpha \wedge \beta))$$
whenever,
- $\alpha,\ \beta \in \Sigma_2$, and
- $\alpha,\ \beta \in \Pi_2$.
- In other words: $\alpha,\ \beta \in \Delta_2$

- Z $\quad (A \rhd B) \wedge (B \rhd A) \rightarrow (A \rhd (A \wedge B)) \quad$ for A and B in ED_2

- Z $\quad (A \rhd B) \wedge (B \rhd A) \rightarrow (A \rhd (A \wedge B))$ for A and B in ED_2
- $$\mathrm{ED}_2 \quad := \quad \Box\mathcal{A} \mid \neg\mathrm{ED}_2 \mid \mathrm{ED}_2 \wedge \mathrm{ED}_2 \mid \mathrm{ED}_2 \vee \mathrm{ED}_2$$

- Z $\quad (A \rhd B) \wedge (B \rhd A) \to (A \rhd (A \wedge B)) \quad$ for A and B in ED_2
- $$\mathrm{ED}_2 \quad := \quad \Box \mathcal{A} \mid \neg \mathrm{ED}_2 \mid \mathrm{ED}_2 \wedge \mathrm{ED}_2 \mid \mathrm{ED}_2 \vee \mathrm{ED}_2$$
- Is this all?

The logic IL

L1: $\Box(A \to B) \to (\Box A \to \Box B)$
L2: $\Box A \to \Box\Box A$
L3: $\Box(\Box A \to A) \to \Box A$

J1: $\Box(A \to B) \to A \rhd B$
J2: $(A \rhd B) \wedge (B \rhd C) \to A \rhd C$
J3: $(A \rhd C) \wedge (B \rhd C) \to A \vee B \rhd C$
J4: $A \rhd B \to (\Diamond A \to \Diamond B)$
J5: $\Diamond A \rhd A$

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.
- R is conversely well-founded and transitive

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.
- R is conversely well-founded and transitive
- $yS_xz \rightarrow xRy \wedge xRz$

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.
- R is conversely well-founded and transitive
- $yS_xz \to xRy \wedge xRz$
- $xRyRz \to yS_xz$

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.
- R is conversely well-founded and transitive
- $yS_xz \rightarrow xRy \wedge xRz$
- $xRyRz \rightarrow yS_xz$
- S_x is transitive and reflexive for each x

- A Veltman frame $F = \langle W, R, S \rangle$,
  $R \subseteq W \times W$,
  $S_w \subseteq W \times W$ for each $w \in W$.
- R is conversely well-founded and transitive
- $yS_xz \rightarrow xRy \wedge xRz$
- $xRyRz \rightarrow yS_xz$
- S_x is transitive and reflexive for each x

A model $M = \langle W, R, S, \Vdash \rangle$,
$\Vdash \subseteq W \times \mathrm{Prop}$

- $w \nVdash \bot$

A model $M = \langle W, R, S, \Vdash \rangle$,
$\Vdash \subseteq W \times \mathrm{Prop}$

- $w \nVdash \bot$
- $w \Vdash A \to B$ iff $w \nVdash A$ or $w \Vdash B$

A model $M = \langle W, R, S, \Vdash \rangle$,
$\Vdash \subseteq W \times \text{Prop}$

- $w \nVdash \bot$
- $w \Vdash A \to B$ iff $w \nVdash A$ or $w \Vdash B$
- $w \Vdash \Box A$ iff $\forall v \ (wRv \Rightarrow v \Vdash A)$

A model $M = \langle W, R, S, \Vdash \rangle$,
$\Vdash \subseteq W \times \mathrm{Prop}$

- $w \nVdash \bot$
- $w \Vdash A \rightarrow B$ iff $w \nVdash A$ or $w \Vdash B$
- $w \Vdash \Box A$ iff $\forall v\ (wRv \Rightarrow v \Vdash A)$
- $w \Vdash A \rhd B$ iff $\forall u\ (wRu \wedge u \Vdash A \Rightarrow \exists v(uS_w v \Vdash B))$

- Montagna has a nice frame condition

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

- Montagna has a nice frame condition

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

- Beklemishev is somewhat similar

A B-simulation on a frame is a binary relation $\mathcal{S}$ for which the following holds.

1. $\mathcal{S}(x, x') \to x{\uparrow} = x'{\uparrow}$
2. $\mathcal{S}(x, x')\ \&\ xRy \to \exists y'(yS_x y' \wedge \mathcal{S}(y, y') \wedge y'S_{x'}{\uparrow} \subseteq yS_x{\uparrow})$

$F \models \mathcal{C}_{\mathrm{B}}$ if and only if there is a B-simulation $\mathcal{S}$ on F such that for all x and y,

$$xRy \to \exists y'(yS_x y' \wedge \mathcal{S}(y, y') \wedge \forall d, e\ (y'S_x dRe \to yRd)).$$

- $$\begin{array}{lcl} \mathrm{ES}_2^0 & := & \mathrm{ED}_2 \\ \mathrm{ES}_2^{n+1} & := & \mathrm{ES}_2^n \mid \mathrm{ES}_2^{n+1} \wedge \mathrm{ES}_2^{n+1} \mid \mathrm{ES}_2^{n+1} \vee \mathrm{ES}_2^{n+1} \mid \\ & & \neg(\mathrm{ES}_2^n \rhd \mathrm{Form}) \end{array}$$

- $$\begin{array}{lll} \mathsf{ES}_2^0 & := & \mathsf{ED}_2 \\ \mathsf{ES}_2^{\mathsf{n}+1} & := & \mathsf{ES}_2^{\mathsf{n}} \mid \mathsf{ES}_2^{\mathsf{n}+1} \wedge \mathsf{ES}_2^{\mathsf{n}+1} \mid \mathsf{ES}_2^{\mathsf{n}+1} \vee \mathsf{ES}_2^{\mathsf{n}+1} \mid \\ & & \neg(\mathsf{ES}_2^{\mathsf{n}} \rhd \mathsf{Form}) \end{array}$$
- $$\begin{array}{lll} \mathcal{S}_0(b,u) & := & b{\uparrow}{=}u{\uparrow} \\ \mathcal{S}_{n+1}(b,u) & := & \mathcal{S}_n(b,u) \wedge \\ & & \forall c\ (bRc \to \exists c'\ (uRc' \wedge \mathcal{S}_n(c,c') \wedge \\ & & cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})) \end{array}$$

- $$\begin{array}{lll} \mathrm{ES}_2^0 & := & \mathrm{ED}_2 \\ \mathrm{ES}_2^{n+1} & := & \mathrm{ES}_2^{n} \mid \mathrm{ES}_2^{n+1} \wedge \mathrm{ES}_2^{n+1} \mid \mathrm{ES}_2^{n+1} \vee \mathrm{ES}_2^{n+1} \mid \\ & & \neg(\mathrm{ES}_2^{n} \rhd \mathrm{Form}) \end{array}$$
- $$\begin{array}{lll} \mathcal{S}_0(b,u) & := & b{\uparrow}{=}u{\uparrow} \\ \mathcal{S}_{n+1}(b,u) & := & \mathcal{S}_n(b,u) \wedge \\ & & \forall c\ (bRc \rightarrow \exists c'\ (uRc' \wedge \mathcal{S}_n(c,c') \wedge \\ & & cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})) \end{array}$$
- For every i we define the frame condition $\mathcal{C}_i$ to be $\forall\, a,b\ (aRb \rightarrow \exists u\ (bS_au \wedge \mathcal{S}_i(b,u) \wedge \forall\, d,e\ (uS_adRe \rightarrow bRe)))$.

- $$\begin{array}{lcl} \mathrm{ES}_2^0 & := & \mathrm{ED}_2 \\ \mathrm{ES}_2^{n+1} & := & \mathrm{ES}_2^{n} \mid \mathrm{ES}_2^{n+1} \wedge \mathrm{ES}_2^{n+1} \mid \mathrm{ES}_2^{n+1} \vee \mathrm{ES}_2^{n+1} \mid \\ & & \neg(\mathrm{ES}_2^{n} \rhd \mathrm{Form}) \end{array}$$
- $$\begin{array}{lcl} \mathcal{S}_0(b,u) & := & b{\uparrow}{=}u{\uparrow} \\ \mathcal{S}_{n+1}(b,u) & := & \mathcal{S}_n(b,u) \wedge \\ & & \forall c\ (bRc \rightarrow \exists c'\ (uRc' \wedge \mathcal{S}_n(c,c') \wedge \\ & & cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})) \end{array}$$
- For every i we define the frame condition $\mathcal{C}_i$ to be $\forall\, a,b\ (aRb \rightarrow \exists u\ (bS_au \wedge \mathcal{S}_i(b,u) \wedge \forall\, d,e\ (uS_adRe \rightarrow bRe)))$.

- $$\begin{array}{lcl} \mathrm{ES}_2^0 & := & \mathrm{ED}_2 \\ \mathrm{ES}_2^{n+1} & := & \mathrm{ES}_2^n \mid \mathrm{ES}_2^{n+1} \wedge \mathrm{ES}_2^{n+1} \mid \mathrm{ES}_2^{n+1} \vee \mathrm{ES}_2^{n+1} \mid \\ & & \neg(\mathrm{ES}_2^n \rhd \mathrm{Form}) \end{array}$$
- $$\begin{array}{lcl} \mathcal{S}_0(b,u) & := & b{\uparrow}{=}u{\uparrow} \\ \mathcal{S}_{n+1}(b,u) & := & \mathcal{S}_n(b,u)\wedge \\ & & \forall c\ (bRc \to \exists c'\ (uRc' \wedge \mathcal{S}_n(c,c')\wedge \\ & & cS_bc' \wedge c'S_u{\uparrow} \subseteq cS_b{\uparrow})) \end{array}$$
- For every i we define the frame condition $\mathcal{C}_i$ to be $\forall\, a, b\ (aRb \to \exists u\ (bS_a u \wedge \mathcal{S}_i(b,u) \wedge \forall\, d, e\ (uS_a dRe \to bRe)))$.

- Theorem
  A finite frame F validates all instances of Beklemishev's principle if and only if $\forall i\ F \models \mathcal{C}_i$.

- $\mathsf{B} \vdash \mathsf{Z}$

- $\mathsf{B} \vdash \mathsf{Z}$
- $\mathsf{B} \models \mathsf{Z}$

- $\mathsf{B} \vdash \mathsf{Z}$
- $\mathsf{B} \models \mathsf{Z}$
- Frame condition Zambella?