Principles of interpretability logic in the intersection of ILP and ILM

Vicent Navarro Arroyo
joint work with
Joost J. Joosten
July 7th-11th
Logic Colloquium 2025, Vienna, Austria

Department of Philosophy,
University of Barcelona

Motivation

Motivation

In Provability Logic, for a fixed theory T (PL) $\Box A$ reads as

$$\text{“}A\text{” is provable in } T.$$

Interpretability Logic (IL) extends PL adding $A \rhd B$ which means

$$T + A \text{ interprets } T + B$$

We say that S interprets T – $S \rhd T$ – if there exists a mapping

$$j\colon \text{Form}_T \to \text{Form}_S$$

that *preserves structure*, for example, if $\circ$ is a binary logical connective, then $(\varphi \circ \psi)^j = \varphi^j \circ \psi^j$ such that moreover

$$\forall \varphi \Big(\Box_T \varphi \to \Box_S \varphi^j \Big).$$

Example

Natural numbers can be interpreted as sets.

Motivation

Gödel's Second Incompleteness Theorem is modally expressed as

$$\Diamond \top \to \neg \Box \Diamond \top.$$

In interpretability logic it can be generalized to

$$\Diamond \top \to \neg(\top \rhd \Diamond \top). \qquad \text{(Feferman)}$$

Motivation

We can define the interpretability logic of a theory T.

$$\mathrm{IL}(T) := \{A \mid \forall * \; T \vdash A^*\},$$

where A is a formula in the language $L_{\Box, \rhd}$

$$F := \bot \mid \mathrm{Prop} \mid F \to F \mid \Box F \mid F \rhd F,$$

and $*$ is a translation sending propositional variables to arithmetical sentences.

Motivation

The axioms of the basic interpretability **IL** are

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 $\boxed{\Box(A \to B) \to A \rhd B}$

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 $\boxed{\Diamond A \rhd A}$

Remark

- J1 tells us that the identity translation yields an interpretation.
- J5 represents Henkin's completeness theorem formalised.

Motivation

There are some interesting principles of interpretability. Namely,

$$\mathsf{M} := A \rhd B \to A \wedge \Box C \rhd B \wedge \Box C \qquad \text{(Montagna)}$$

$$\mathsf{P} := A \rhd B \to \Box(A \rhd B) \qquad \text{(Persistence)}$$

It is known that

$$\mathsf{IL}(\mathsf{PA}) := \mathsf{IL}M \quad \text{(Full induction)}$$

and

$$\mathsf{IL}(I\Sigma_1) := \mathsf{ILP} \quad \text{(Finitely Axiomatized).}$$

ILM and ILP motivate the characterisation of IL(All).

$$\mathrm{IL(All)} := \{A \mid \forall T \supseteq I\Delta_0 + \mathrm{Exp}\ \forall * \ T \vdash A^*\},$$

the interpretability logic of al "reasonable" arithmetical theories.

Remark

$\mathrm{IL(All)} \subsetneq \mathrm{ILM} \cap \mathrm{ILP}$

We present some advances on its modal characterization.

Semantics and intersections

Semantics and intersections

In interpretability logic, models are 4-tuples

$$\mathcal{M} := \langle W, R, \{S_x\}_{x \in W}, V \rangle$$

where

- $W \neq \varnothing$
- $R \subseteq W \times W$
- $S_x \subseteq x\upharpoonright \times x\upharpoonright$
- $V\colon \mathrm{Prop} \to \mathcal{P}(W)$

$x\upharpoonright := \{y \mid xRy\}$.

R transitive and conversely well-founded;

S_x is reflexive transitive and contains R on $x\upharpoonright$.

$\mathcal{F} = \langle W, R, \{S_x\}_{x \in W} \rangle$ denotes a frame.

Sometimes we denote models as $\mathcal{M} = \langle \mathcal{F}, V \rangle$.

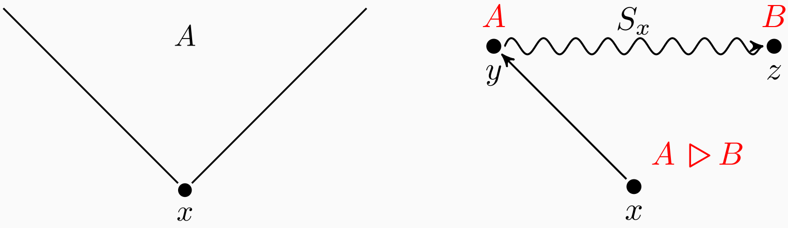
Propositions, implications and *falsum* ($\bot$) are forced as usual.

The forcing of formulas $\Box A$ is

$$\mathcal{M}, x \Vdash \Box A :\iff \forall y (xRy \to \mathcal{M}, y \Vdash A).$$

The forcing of formulas $A \rhd B$ is

$$\mathcal{M}, x \Vdash A \rhd B :\iff \forall y (xRy \land \mathcal{M}, y \Vdash A \to \exists z \colon yS_x z \land \mathcal{M}, z \Vdash B).$$



(a) (b)

Figure 1: (a) $\Box A$ is forced at x (b) $A \rhd B$ is forced at x

Semantics and intersection

Validity on models and frames is defined as follows.

Validity

- **Validity of a formula on a model:**

$$\mathcal{M} \vDash \varphi \text{ iff } \mathcal{M}, w \Vdash \varphi, \text{ for all } w \in W.$$

- **Validity of a formula on a frame:**

$$\mathcal{F} \vDash \varphi \text{ iff } \forall V \langle \mathcal{F}, V \rangle \vDash \varphi.$$

- **Validity of a scheme:** A model or a frame validates a scheme X ($\mathcal{M} \vDash$ X and $\mathcal{F} \vDash$ X, respectively) iff it validates all X's instances.

Semantics and intersections

The **frame condition** of a scheme X is a first (or higher) order predicate formula $\mathcal{C}$ such that

$$\forall \mathcal{F}(\mathcal{F} \models \mathcal{C} \iff \mathcal{F} \models \mathsf{X}).$$

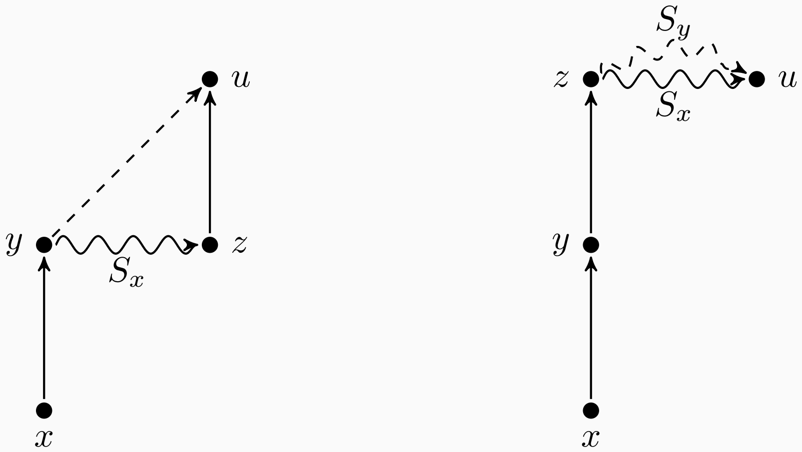
Example

$$\mathcal{F} \models \Box A \to \Box\Box A \iff \mathcal{F} \models \forall x, y, z \left(xRy \wedge yRz \to xRz\right)$$

Frame conditions of ILM and ILP.

$$\mathcal{F} \models \mathsf{M} \iff \mathcal{F} \models xRyS_x zRu \to yRu.$$

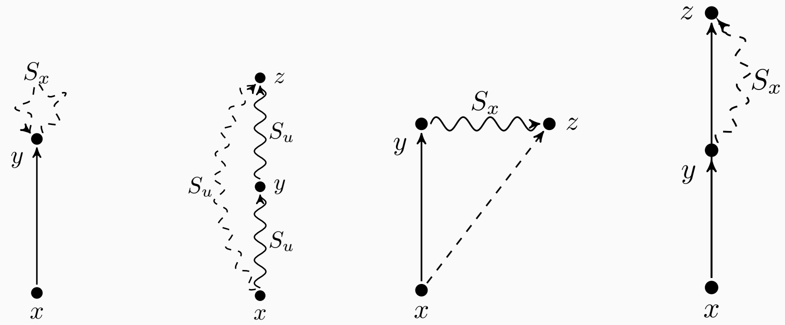
$$\mathcal{F} \models \mathsf{P} \iff \mathcal{F} \models xRyRzS_x u \to zS_y u.$$



(a) (b)

Figure 2: Frame condition of M (a) Frame condition of P (b)

Semantics and intersections



(a) J1 (b) J2 (c) J4 (d) J5

Figure 3: Frame definition reflecting axioms $\Box(A \to B) \to A \rhd B$ (J1), $A \rhd B \wedge B \rhd C \to A \rhd C$ (J2), $A \rhd B \to (\Diamond A \to \Diamond B)$ (J4) and $\Diamond A \rhd A$ (J5)

Semantics and intersections

Sometimes we need to close on the frame properties.

Closure

The closure of a (proto-) frame $\mathcal{F} := \langle W, R, \{S_x\}_{x \in W} \rangle$ under some principle X is the smallest structure $\overline{\mathcal{F}}^X := \langle W, \overline{R}^X, \{\overline{S}_x^X\}_{x \in W} \rangle$ satisfying X such that $R \subseteq \overline{R}^X$ and $S_x \subseteq \overline{S}_x^X$, for every $x \in W$.

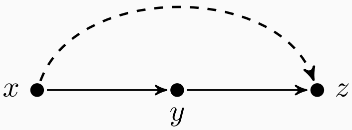


Figure 4: Transitive closure

Semantics and intersection

Frame operator

If $L = \{\phi_i\}_i$ is a set of atomic predicates (like xRy or yS_xz, etc.), we define the **IL-frame induced by** L, $\overline{\mathcal{F}(\bigwedge_i \phi_i)}^{\mathsf{IL}}$, as the universal closure of the smallest proto-frame that satisfies all atomic predicates.

For brevity, we will write $\mathcal{F}(\bigwedge_i \phi_i)$.

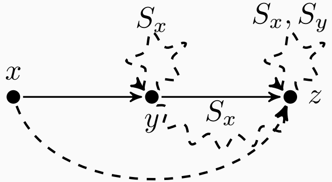


Figure 5: Closure of $\{xRy, yRz\}$ under **IL** frame requirements.

Semantics and intersection

Let $\mathfrak{F}$ be a class of IL-frames. We define the interpretability logic corresponding to $\mathfrak{F}$.

$$\mathsf{IL}[\mathfrak{F}] := \{A \colon \text{for all } \mathcal{F} \in \mathfrak{F},\ \mathcal{F} \models A\}.$$

Let $F(x, y, z)$ denote any first or higher order formula where the only free variables are x, y, z. We now define the following class of conditions.

$$\mathcal{C}_{\mathsf{ILP} \cap_{\mathsf{S}} \mathsf{ILM}} :=$$
$$\{F(x, y, z) \to xS_y z \colon \mathsf{ILP} \models F(x, y, z) \to xS_y z \ \wedge\ \mathsf{ILM} \models F(x, y, z) \to xS_y z\}.$$

Also, we define the class

$$\mathfrak{All} := \{\mathcal{F} \models \mathsf{ILW} \colon \forall C \in \mathcal{C}_{\mathsf{ILP} \cap_{\mathsf{S}} \mathsf{ILM}}, \mathcal{F} \models C\}.$$

The principle W is

$$\mathsf{W} := A \rhd B \to A \rhd (B \land \Box \neg A)$$

and its frame condition is that there are no $S_x; R$ infinite chains.

Conjecture 1 (Goris, Joosten 2020)

$$\mathbf{IL}(\mathrm{All}) = \mathbf{IL}[\mathfrak{All}].$$

Recall

$$\mathrm{IL}(\mathrm{All}) := \{A \mid \forall T \supseteq I\Delta_0 + \mathrm{Exp}\ \forall * \ T \vdash A^*\}.$$

Semantics and intersection

M ∩ P-closure

Given a proto-frame $\mathcal{F} = \langle W, R, S \rangle$, its M ∩ P-**closure** is $\overline{\mathcal{F}}^{\mathsf{M} \cap \mathsf{P}} := \overline{\mathcal{F}}^{\mathsf{M}} \cap \overline{\mathcal{F}}^{\mathsf{P}} = \langle W, \overline{R}^{\mathsf{M}} \cap \overline{R}^{\mathsf{P}}, \overline{S}^{\mathsf{M}} \cap \overline{S}^{\mathsf{P}} \rangle$.

As an example, consider the principle M_0

$$\mathsf{M}_0 := A \rhd B \to \Diamond A \land \Box C \rhd B \land \Box C,$$

whose frame condition is

$$\forall x, y, z, u, v \Big(xRyRzS_x uRv \to yRv \Big).$$

Semantics and intersection

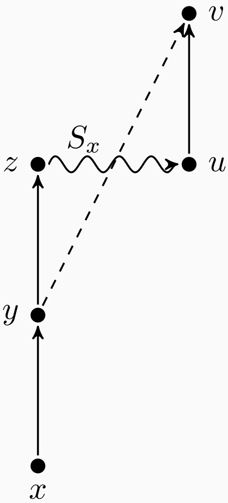
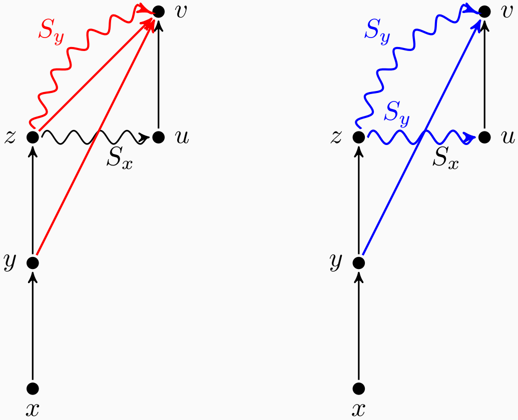


Figure 6: M_0

Semantics and intersection



(a) (b)

Figure 7: (a) M closure and (b) P closure

Semantics and intersection

$\mathsf{M} \bigcap_{\mathcal{F}} \mathsf{P}$-clause set

We define the $\mathsf{M} \cap_{\mathcal{F}} \mathsf{P}$-clause set as

$$\bigwedge_i \phi_i \to \varphi :\in \mathsf{M} \cap_{\mathcal{F}} \mathsf{P} \text{ iff } \overline{\mathcal{F}(\bigwedge_i \phi_i)}^{\mathsf{M} \cap \mathsf{P}} \vDash \varphi$$

whenever $\{\phi_i\}_i \cup \{\varphi\}$ is a set of atomic predicates so that $\mathcal{F}(\bigwedge_i \phi_i)$ defines a proto-frame.

Remark

$\bigwedge_i \phi_i \to \varphi$ is a Horn clause.

Non-empty since the M_0 frame condition belongs to it.

It is known that the *Broad* series and the *Slim* hierarchy belong to it.

Semantics and intersection

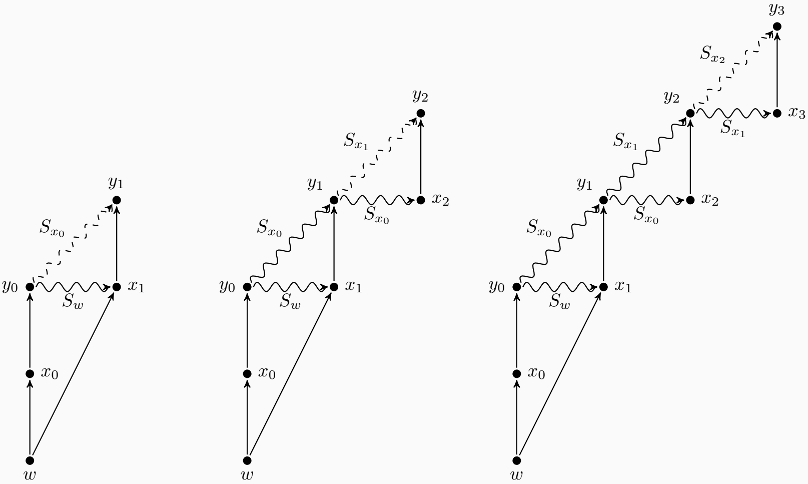


Figure 8: Slim (or Staircase) hierarchy

Semantics and intersection

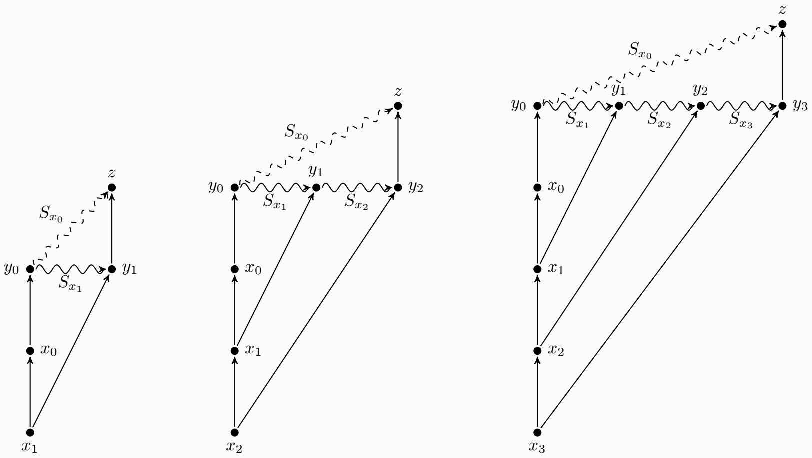


Figure 9: Broad series

Semantics and intersection

$\mathsf{M} \cap_{\mathcal{F}} \mathsf{P}$ defines a fragment of $\mathbf{IL}[\mathfrak{All}]$.

Let us define the lower-case class of $\mathbf{IL}$-frames

$$\mathfrak{all} := \{\mathcal{F} \models \mathbf{IL}\mathsf{W} \colon \forall C \in \mathsf{M} \cap_{\mathcal{F}} \mathsf{P},\ \mathcal{F} \models C\}.$$

Theorem

$$\mathbf{IL}[\mathfrak{all}] \ \subseteq\ \mathbf{IL}[\mathfrak{All}].$$

Remark

- It is unknown if $\mathbf{IL}[\mathfrak{all}] \ \subset\ \mathbf{IL}[\mathfrak{All}]$.
- $\mathbf{IL}[\mathfrak{all}]$ entails the frame conditions of *Broad* and *Slim*.

It is natural to conjecture that

Conjecture 2

$$\mathsf{IL}[\mathfrak{all}] = \mathsf{IL}(\mathsf{All}).$$

This new conjecture strengthens the old conjecture.

Conjecture 1 (Goris, Joosten 2020)

$$\mathsf{IL}(\mathsf{All}) = \mathsf{IL}[\mathfrak{All}].$$

How can we get a grip on $\mathsf{M} \bigcap_{\mathcal{F}} \mathsf{P}$?

One may try to focus on the clauses that imply an *R*-pair and conjecture that

Conjecture 3

Consider an **IL**-frame $\mathcal{F} = \langle W, R, S \rangle$. Then, for any $x, y \in W$, we have that $x\overline{R}^{\mathsf{M}}y \wedge x\overline{R}^{\mathsf{P}}y \wedge \neg(xRy) \to x\overline{R}^{\mathsf{M}_0}y$.

Nonetheless, this is disproven by the...

Pencil frame

Pencil frame

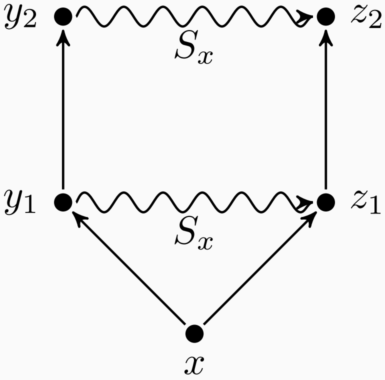


Figure 10: Pencil frame.

Pencil frame

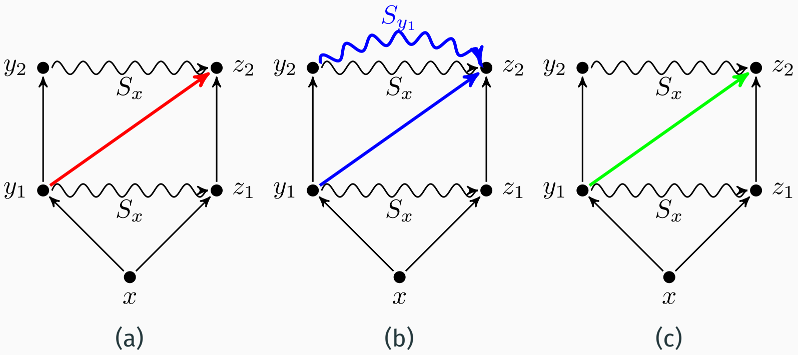


Figure 11: (a) M-closure (b) P-closure (c) Intersection.

Remark

Observe the green arrow is not in the M_0-closure.

Pencil frame

We observe the Pencil frame is **not** modally definable.

Frame definability

Given a first or higher order predicate formula $\mathcal{C}$. The class of frames that make true $\mathcal{C}$ is **modally definable** if

$$\exists A \in L_{\Box,\rhd}\, \forall \mathcal{F}\, (\mathcal{F} \vDash \mathcal{C} \iff \mathcal{F} \vDash A).$$

Example

The class of transitive frames is defined by $\Box A \to \Box\Box A$.

Remark

Consider the formula $\mathcal{C}_P := xRy_1S_xz_1Rz_2 \wedge y_1Ry_2S_xz_2 \to y_1Rz_2$.
Notice that y_1Rz_2 is precisely the green arrow.

Pencil frame

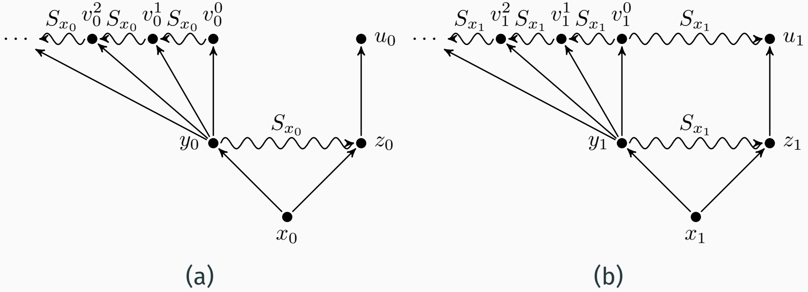


Figure 12: (a) $\mathcal{F}_0$ satisfies $\mathcal{C}_P$ (b) $\mathcal{F}_1$ does not satisfy $\mathcal{C}_P$

Pencil frame

Theorem: Pencil frame is not modally definable

- By *Reductio ad Absurdum*, assume it is, that is,

$$\exists A \in L_{\Box,\rhd} \forall \mathcal{F} \, (\mathcal{F} \vDash \mathcal{C}_P \iff \mathcal{F} \vDash A).$$

- Consider $\mathcal{F}_0$ and $\mathcal{F}_1$. Notice $\mathcal{F}_0 \vDash \mathcal{C}_P$ whereas $\mathcal{F}_1 \nvDash \mathcal{C}_P$.
- Then, by hypothesis, $\mathcal{F}_0 \vDash A$ and $\mathcal{F}_1 \nvDash A$.
- **Claim:** $\forall V_1 \exists V_0 : \langle \mathcal{F}_1, V_1 \rangle \sim_{bisimilar} \langle \mathcal{F}_0, V_0 \rangle$.
- Bisimilar image-finite models prove the same modal formulas (**Hennessy–Milner**).
- Thus, $\exists V_0 \langle F_0, V_0 \rangle \nvDash A$. Contradiction! ($\mathcal{F}_0 \vDash A$)

New series

New series

Given that the Pencil frame is not modally definable and its frame condition is in $\mathsf{M}\bigcap_{\mathcal{F}}\mathsf{P}$ and not induced by neither *Broad* nor *Slim*, a natural question arises:

Is there a class of modally definable frames whose frame condition is in $\mathsf{M}\bigcap_{\mathcal{F}}\mathsf{P}$ but it is not induced by *Slim* nor *Broad*?

We found out that the answer is positive 😄

New series

We will inductively define a series of schemes.

Firstly, we inductively define the following series of formulas.

$$\varphi^0 := \Diamond\left((D \rhd D_0) \wedge \Diamond\neg(A \rhd \neg C)\right),$$
$$\varphi^n := \Diamond\left((D_{n-2} \rhd D_{n-1}) \wedge \varphi^{n-1}\right). \qquad (n \geq 1)$$

Then, we inductively define V as the series of all the principles V^n, for any $n \in \mathbb{N}$, where

$$\begin{aligned}
\mathsf{V}^0 &:= A \rhd B \rightarrow \Big((D_0 \rhd \Diamond D_1) \wedge \varphi^0\Big) \rhd B \wedge \Box C \wedge (D \rhd D_1),\\
\mathsf{V}^{n+1} &:= \mathsf{V}^n[\varphi^n/\varphi^{n+1};\\
&\qquad D_n \rhd \Diamond D_{n+1}/D_{n+1} \rhd \Diamond D_{n+2};\\
&\qquad D \rhd D_{n+1}/D \rhd D_{n+2}]
\end{aligned}$$

New series

For example,

$$\mathsf{V}^0 := A \rhd B \to \left((D_0 \rhd \Diamond D_1) \wedge \Diamond \left((D \rhd D_0) \wedge \Diamond \neg(A \rhd \neg C)\right)\right) \rhd B \wedge C \wedge (D \rhd D_1),$$

$$\mathsf{V}^1 := A \rhd B \to \Big((D_1 \rhd \Diamond D_2) \wedge \Diamond \left((D_0 \rhd D_1) \wedge \Diamond \left((D \rhd D_0) \wedge \Diamond \neg(A \rhd \neg C)\right)\right)\Big) \rhd B \wedge C \wedge (D \rhd D_2).$$

Their frame conditions are, respectively, . . .

New series

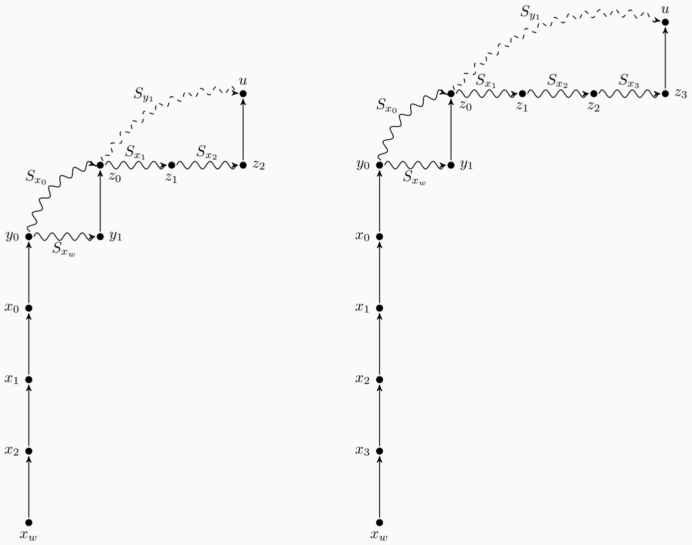
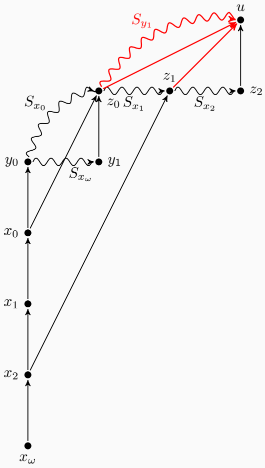
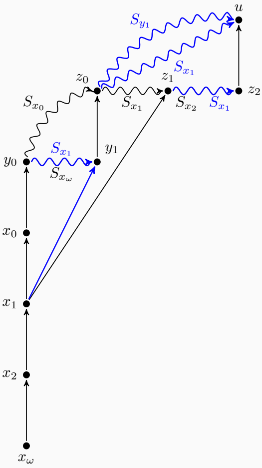


Figure 13: (Left) Frame condition of V_0. (Right) Frame condition of V_1

New series



(a)



(b)

Figure 14: (a) M-closure (b) P-closure

New series

Remark

The V series is **not** a hierarchy.

Wrapping up:

- The classes of frames that satisfy the frame conditions of the V series are modally definable.
- The frame conditions of the V series belong to $\mathrm{M}\bigcap_{\mathcal{F}}\mathrm{P}$.
- It can be shown that neither the *Broad* series nor the *Slim* hierarchy induce the V series.

Also, these principles of the V series are arithmetically valid through **arithmetical definable cuts**.

Summary (Summer-e)

1. We strengthen the old conjecture by focusing on Horn clauses;
2. We show that all known principles fall in this class;
3. We show that some frame properties are modally undefinable;
4. We found a new series of principles;
5. We have proven the new principles to be arithmetically sound;
6. Thus the conjecture still stands;
7. Preprint and paper coming out ‘soon’.

Thank you! Danke!