Epistemic Arithmetic

Eric Pacuit

University of Maryland

Lecture 5, ESSLLI 2025

August 1, 2025

Plan

- ✓ Introduction: Smullyan's Machine
- ✓ Background
  - ✓ Formal Arithmetic
  - ✓ Gödel's Incompleteness Theorems
  - ✓ Names and Gödel numbering
  - ✓ Fixed Point Theorem
- ✓ Provability predicate and Löb's Theorem
- ✓ Provability logic
- ✓ Predicate approach to modality
- ✓ The Knower Paradox and variants
- ✓ A Primer on Epistemic and Doxastic Logic
- ✓ Anti-Expert Paradox, and related paradoxes
- ▶ Predicate approach to modality, continued
- ▶ Epistemic Arithmetic
- ▶ Gödel's Disjunction

Operator > Predicate

✓ Montague provided the first result by proving that the predicate version of the modal system **T** is inconsistent if it is combined with weak systems of arithmetic. From his result he concluded that "virtually all of modal logic...must be sacrificed", if necessity is conceived of as a predicate of sentences.

⇒ The other technical achievement that brought about the triumph of the operator view was the emergence of possible-worlds semantic. Hintikka, Kanger and Kripke provided semantics for modal operator logics, while nothing similar seemed available for the predicate approach.

Volker Halbach, Hannes Leitgeb and Philip Welch (2003). *Possible-Worlds Semantics for Modal Notions Conceived as Predicates*. Journal of Philosophical Logic, 32:2, pp. 179-223.

A **frame** is a tuple (W, R) where W is a nonempty set and R is a relation on W.

A **frame** is a tuple (W, R) where W is a nonempty set and R is a relation on W.

A **PW-model** is a triple (W, R, V) such that (W, R) is a frame and V assigns to every $w \in W$ as subset of $\mathcal{L}_\Box$ such that:

$$V(w) = \{A \in \mathcal{L}_\Box \mid \text{ for all } u, \text{ if } w\,R\,u, \text{ then } V(u) \models A\}$$

A **frame** is a tuple (W, R) where W is a nonempty set and R is a relation on W.

A **PW-model** is a triple (W, R, V) such that (W, R) is a frame and V assigns to every $w \in W$ as subset of $\mathcal{L}_\Box$ such that:

$$V(w) = \{A \in \mathcal{L}_\Box \mid \text{ for all } u, \text{ if } w\, R\, u, \text{ then } V(u) \models A\}$$

If (W, R, V) is a model, we say that the frame (W, R) **supports** the model (W, R, V) or that (W, R, V) is **based on** (W, R).

A frame **admits a valuation** if there is a valuation V such that (W, R, V) is model.

$V(w) \models \Box\ulcorner A \urcorner$ iff for all $v \in W$, if $w\, R\, v$, then $V(v) \models A$

$$V(w) \models \Box\ulcorner A \urcorner \text{ iff for all } v \in W\text{, if } w \, R \, v\text{, then } V(v) \models A$$

Characterization Problem: Which frames support PW-models?

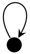
$$V(w) \models \Box\ulcorner A \urcorner \text{ iff for all } v \in W, \text{ if } w\, R\, v, \text{ then } V(v) \models A$$

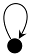
Characterization Problem: Which frames support PW-models?

Lemma (Normality). Suppose (W, R, V) is a PW-model, $w \in W$ and $A, B \in \mathcal{L}_\Box$. Then the following holds:

- If $V(u) \models A$ for all $u \in W$, then $V(w) \models \Box\ulcorner A \urcorner$.
- $V(w) \models \Box(\ulcorner A \to B \urcorner) \to (\Box\ulcorner A \urcorner \to \Box\ulcorner B \urcorner)$

$$\forall x \forall y ((\mathsf{Sent}(x) \land \mathsf{Sent}(y)) \to (\Box \ulcorner x \underset{\cdot}{\to} y \urcorner \to (\Box x \to \Box y)))$$





Fact (Tarski). The above frame with one world that sees itself does not admit a valuation.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A \urcorner$, and so it holds at every world.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A\urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A\urcorner$.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A\urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A\urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A\urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A\urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.
- Thus, $V(w) \models A$.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

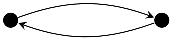
- We have **PA** $\vdash A \leftrightarrow \neg\Box\ulcorner A\urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A\urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.
- Thus, $V(w) \models A$.
- Hence, $V(w) \models \neg\Box\ulcorner A\urcorner$; and so, there is some u such that $w\,R\,u$ and $V(u) \models \neg A$.

Fact (Montague's Theorem). If (W, R) admits a valuation, then (W, R) is not reflexive.

Assume (W, R, V) is a PW-model based on (W, R) which is reflexive.

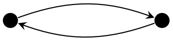
- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A \urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A \urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.
- Thus, $V(w) \models A$.
- Hence, $V(w) \models \neg\Box\ulcorner A \urcorner$; and so, there is some u such that $w\,R\,u$ and $V(u) \models \neg A$.
- Again, using the same argument as above, $V(u) \models A$. Contradiction.

1. The following frame does not admit a valuation:



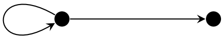
   Use the fixed point: $A \leftrightarrow \neg\Box\ulcorner\Box\ulcorner A\urcorner\urcorner$

1. The following frame does not admit a valuation:

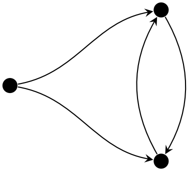
   Use the fixed point: $A \leftrightarrow \neg\Box\ulcorner\Box\ulcorner A\urcorner\urcorner$

2. The following frame does not admit a valuation:

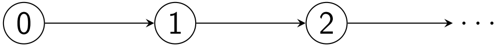
   Use the fixed point: $A \leftrightarrow (\Box\ulcorner A\urcorner \to \Box\ulcorner\neg A\urcorner)$

3. The following frame does not admit a valuation:



Use the fixed point: $A \leftrightarrow (\neg\Box\ulcorner\Box\ulcorner A\urcorner\urcorner \land \neg\Box\ulcorner A\urcorner)$

4. The following frame $(\mathbb{N}, succ)$ does not admit a valuation:



Use the fixed point: $A \leftrightarrow \neg \forall x \Box \underset{\cdot}{h}(x, \ulcorner A \urcorner)$

where $\underset{\cdot}{h}$ represents a function that applies n-boxes to B:

$$h(n) = \ulcorner \Box \cdots \ulcorner \Box \ulcorner B \urcorner \urcorner \cdots \urcorner$$

V. McGee (1985). *How truthlike can a predicate be? A negative result*. Journal of Philosophical Logic, 14, pp. 399-410.

A. Visser (1989). *Semantics and the Liar paradox*. in Handbook of Philosophical Logic, Vol. 4, Reidel, Dordrecht.

Lemma. Let (W, R, V) be a PW-model based on a transitive frame. Then,

$$\Box\ulcorner A\urcorner \rightarrow \Box\ulcorner\Box\ulcorner A\urcorner\urcorner$$

obtains for all $w \in W$ and sentences $A \in \mathcal{L}_\Box$.

Löb's Theorem For every world w in a PW-model based on a transitive frame and every sentence $A \in \mathcal{L}_\Box$, the following holds:

$$\Box(\ulcorner\Box\ulcorner A\urcorner \rightarrow A\urcorner) \rightarrow \Box\ulcorner A\urcorner$$

Fact. In a transitive frame admitting a valuation every world is either a dead end state or it can see a dead end state.

Fact. In a transitive frame admitting a valuation every world is either a dead end state or it can see a dead end state.

Proof. Since the frame is transitive, Löb's Theorem holds.

Applying Löb's Theorem to $\bot$, we obtain:

$$V(w) \models \Box\ulcorner\bot\urcorner \vee \Diamond\ulcorner\Box\ulcorner\bot\urcorner\urcorner$$

- It is not hard to show that all converse wellfounded frames support a PW-model:

  If (W, R) is converse wellfounded, then define a valuation for (W, R) by induction along R in the following way:

$$V(w) = \{A \in \mathcal{L}_\Box \mid \forall v(w\,R\,v \Rightarrow V(v) \models A\}$$

N. Belnap and A. Gupta (1993). *The Revision Theory of Truth*. The MIT Press.

- It is not hard to show that all converse wellfounded frames support a PW-model:

  If (W, R) is converse wellfounded, then define a valuation for (W, R) by induction along R in the following way:

$$V(w) = \{A \in \mathcal{L}_\Box \mid \forall v(w\, R\, v \Rightarrow V(v) \models A\}$$

N. Belnap and A. Gupta (1993). *The Revision Theory of Truth*. The MIT Press.

- However, there are some converse illfounded frames that admit valuations. Because of these frames the Characterisation Problem is nontrivial.

Predicate Approaches to Modality

Johannes Stern (2016). *Toward Predicate Approaches to Modality*. Springer.

Plan

- ✓ Introduction: Smullyan’s Machine
- ✓ Background
  - ✓ Formal Arithmetic
  - ✓ Gödel’s Incompleteness Theorems
  - ✓ Names and Gödel numbering
  - ✓ Fixed Point Theorem
- ✓ Provability predicate and Löb’s Theorem
- ✓ Provability logic
- ✓ Predicate approach to modality
- ✓ The Knower Paradox and variants
- ✓ A Primer on Epistemic and Doxastic Logic
- ✓ Anti-Expert Paradox, and related paradoxes
- ✓ Predicate approach to modality, continued
- ▶ Epistemic Arithmetic
- ▶ Gödel’s Disjunction

The Incompleteness Theorems

Theorem (Gödel's First Incompleteness Theorem)

Assume that **PA** is Σ^0_1-sound. Then there is a Π^0_1-sentence φ such that **PA** neither proves φ nor $\neg\varphi$.

Theorem (Gödel's Second Incompleteness Theorem)

Assume that **PA** is consistent. Then **PA** cannot prove $\mathrm{Con}_{\mathbf{PA}}$.

$\mathrm{Con}_{\mathbf{PA}}$ *is a* Π^0_1*-statement that informally asserts "for all* x*,* x *does not code a proof of a contradiction from the axioms of* **PA***"*

Do the incompleteness theorems imply that “the mathematical outputs of the idealized human mind do not coincide with the mathematical outputs of any idealized finite machine (Turing machine)”?

Peter Koellner (2016). *Gödel’s Disjunction*. in *Gödel’s Disjunction: The scope and limits of mathematical knowledge*, pp. 148-188, Oxford University Press.

Relative Provability; Absolute Provability; Truth

F an arbitrary formal system with the feature that each sentence of F is true and the rules of F are truth preserving

K the set of all sentences that are "absolutely provable"

T the set of all sentences that are true

Claim 1: For any formal system F, $F \subseteq T \Rightarrow F \subsetneq T$

Claim 2: For any formal system F, $K(F \subseteq T) \Rightarrow F \subsetneq K$

Gödel did *not* conclude that for F, $F \subseteq T \rightarrow F \subsetneq K$

Gödel did *not* conclude that for F, $F \subseteq T \to F \subsetneq K$

Does incompleteness imply that there are **absolutely undecidable** sentences?

> “The statements are not all absolutely undecidable; rather, one can always pass to a “higher” system in which the sentence in question is decidable...Perhaps there is a “master system,” F^* such that relative provability with regard to F^* coincides with absolute provability....What we can conclude is merely that *if* there is a such a master system, then we could never know (in the sense of being able to absolutely prove) that all of its axioms were true.”

Gödel did *not* conclude that for F, $F \subseteq T \rightarrow F \subsetneq K$

Does incompleteness imply that there are **absolutely undecidable** sentences?

> "The statements are not all absolutely undecidable; rather, one can always pass to a "higher" system in which the sentence in question is decidable...Perhaps there is a "master system," F^* such that relative provability with regard to F^* coincides with absolute provability....What we can conclude is merely that *if* there is a such a master system, then we could never know (in the sense of being able to absolutely prove) that all of its axioms were true."

if there is an F such that $F = K$, *then* $K \subsetneq T$

Gödel's Disjunction

Either $(\neg\exists F(F = K))$ or $(\exists\varphi(T(\varphi) \wedge \neg K(\varphi) \wedge \neg K(\neg\varphi)))$

"So the following disjunctive conclusion is inevitable: Either mathematics is incompletable...that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are strictly speaking, three alternatives)."

To make the above arguments precise, we need to spell out the background assumptions on F, K and T.

To make the above arguments precise, we need to spell out the background assumptions on F, K and T.

- Turing provides a substantive analysis of F.

To make the above arguments precise, we need to spell out the background assumptions on F, K and T.

- Turing provides a substantive analysis of F.
- Tarski gives a structural analysis of T.

To make the above arguments precise, we need to spell out the background assumptions on F, K and T.

- Turing provides a substantive analysis of F.
- Tarski gives a structural analysis of T.
- What about K?

  In the case of K there is no hope of giving a substantive analysis; the most that one could hope for is a structural analysis. The trouble is that there is little agreement on the element of idealization involved in the notion of “absolute provability” (i.e., “the idealized human mind”).

Epistemic Arithmetic

P. Koellner (2016). *Gödel's Disjunction*. in Gödel's Disjunction: The Scope and Limit and Mathematical Knowledge, Oxford University Press.

W. Reinhardt (1985). *Absolute Versions of Incompleteness Theorems*. Nous, 19(3), pp. 317 - 346.

W. Reinhardt (1986). *Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems*. Journal of Philosophical Logic, 15, pp. 427 - 474.

The language of **EA** is the language $\mathcal{L}_A$ augmented with a unary sentential operator 'K'.

The axioms of **EA** fall into two categories—the axioms of arithmetic and the axioms of absolute provability.

The language of **EA** is the language $\mathcal{L}_A$ augmented with a unary sentential operator 'K'.

The axioms of **EA** fall into two categories—the axioms of arithmetic and the axioms of absolute provability.

The axioms of arithmetic are simply those of **PA**, only now the induction scheme is taken to hold for all formulas in $\mathcal{L}_{EA}$.

Axioms of absolute provability

E1. Universal closures of formulas of the form $K\varphi$ where φ is a first-order validity.

E2. Universal closures of formulas of the form

$$(K(\varphi \to \psi) \land K\varphi) \to K\psi$$

E3. Universal closures of formulas of the form

$$K\varphi \to \varphi$$

E4. Universal closures of formulas of the form

$$K\varphi \to KK\varphi$$

For a collection Σ of formulas in $\mathcal{L}_{EA}$ we use $K\Sigma$ to denote the collection of formulas '$K\varphi$' where $\varphi \in \Sigma$.

The system **EA** is the theory axiomatized by $\Sigma \cup K\Sigma$ where Σ consists of the axioms of **PA** (in the language $\mathcal{L}_{EA}$) and the basic axioms of absolute provability.

Feferman Dot Notation, I

Arithmetic operations on Gödel numbers is denoted by placing a dot under the associated syntactic symbol.

Feferman Dot Notation, I

Arithmetic operations on Gödel numbers is denoted by placing a dot under the associated syntactic symbol.

For example, $\underset{\cdot}{\neg}$ is the operation where:

$$\underset{\cdot}{\neg}\ulcorner\varphi\urcorner \equiv \ulcorner\neg\varphi\urcorner$$

Similarly for $\underset{\cdot}{\vee}$, $\underset{\cdot}{\rightarrow}$, etc.

Feferman Dot Notation, II

Although it makes sense to write $\mathrm{Prov}_{\mathbf{PA}}(\ulcorner \varphi \urcorner)$, it does not make sense to write $\forall x\, \mathrm{Prov}_{\mathbf{PA}}(x)$.

Feferman Dot Notation, II

Although it makes sense to write $\mathsf{Prov}_{\mathbf{PA}}(\ulcorner\varphi\urcorner)$, it does not make sense to write $\forall x\, \mathsf{Prov}_{\mathbf{PA}}(x)$.

$\forall x\, \mathsf{Prov}_{\mathbf{PA}}(\dot{\overline{x}})$ means that "for every natural number x, if you take the canonical numeral for x, substitute it for the dot in '$\varphi(\cdot)$', then the Gödel number of the resulting expression is in the range of the arithmetical relation 'Prov$_{\mathbf{PA}}$'."

Feferman Dot Notation, III

if x is the Gödel number of a formula, z is the Gödel number of a variable, and y is a natural number, then

$$x(\dot{\bar{y}}/z))$$

is the Gödel number of the formula obtained by

substituting the canonical numeral for y for (the variable numbered by) z in (the expression numbered by) x.

Typed Truth

The language $\mathcal{L}_{EAT}$ is the language $\mathcal{L}_{EA}$ augmented with a unary predicate 'T'.

The predicate 'T' is intended to be a (Tarskian) truth predicate that applies to sentences of the sublanguage $\mathcal{L}_{EA}$ where 'T' is omitted.

Hence we are dealing here with a *typed truth predicate*.

Typed Truth

Define a valuation function *Val* on closed terms:

V1. $\forall x[Val(\dot{\overline{x}}) = x]$

V2. $$\forall x_1 \cdots \forall x_n[(CTerm(x_1) \wedge \cdots \wedge CTerm(x_n)) \\ \rightarrow (Val(f(x_1, \ldots, x_n)) = f(Val(x_1), \ldots, Val(x_n))]$$

Typed Truth

T1. $\forall x[T(x) \rightarrow Sent(x)]$

Typed Truth

T1. $\forall x[T(x) \rightarrow Sent(x)]$

T2. For each atomic formula $R(x_1, \ldots, x_n)$:
$\forall x_1 \cdots \forall x_n[CTerm(x_1) \wedge \cdots \wedge CTerm(x_n)) \rightarrow$
$T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \rightarrow R(Val(x_1), \ldots, Val(x_n))$

Typed Truth

T1. $\forall x[T(x) \rightarrow Sent(x)]$

T2. For each atomic formula $R(x_1, \ldots, x_n)$:
$\forall x_1 \cdots \forall x_n[CTerm(x_1) \wedge \cdots \wedge CTerm(x_n)) \rightarrow$
$T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \rightarrow R(Val(x_1), \ldots, Val(x_n))$

T3. $\forall x[Sent(x) \rightarrow T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x)]$

T4. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)))]$

T5. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow (T(x) \vee T(y)))]$

Typed Truth

T1. $\forall x[T(x) \rightarrow Sent(x)]$

T2. For each atomic formula $R(x_1, \ldots, x_n)$:
$\forall x_1 \cdots \forall x_n[CTerm(x_1) \wedge \cdots \wedge CTerm(x_n)) \rightarrow$
$T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \rightarrow R(Val(x_1), \ldots, Val(x_n))$

T3. $\forall x[Sent(x) \rightarrow T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x)]$

T4. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)))]$

T5. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow (T(x) \vee T(y)))]$

T6. $\forall x \forall z[(Var(z) \wedge Sent((\underset{\cdot}{\forall} z)x)) \rightarrow (T((\underset{\cdot}{\forall} z)x) \leftrightarrow \forall y T(x(\dot{\overline{y}}/z)))]$

Typed Truth

T1. $\forall x[T(x) \rightarrow Sent(x)]$

T2. For each atomic formula $R(x_1, \ldots, x_n)$:
$\forall x_1 \cdots \forall x_n[CTerm(x_1) \wedge \cdots \wedge CTerm(x_n)) \rightarrow$
$T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \rightarrow R(Val(x_1), \ldots, Val(x_n))$

T3. $\forall x[Sent(x) \rightarrow T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x)]$

T4. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)))]$

T5. $\forall x \forall y[(Sent(x) \wedge Sent(y)) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow (T(x) \vee T(y)))]$

T6. $\forall x \forall z[(Var(z) \wedge Sent((\underset{\cdot}{\forall} z)x)) \rightarrow (T((\underset{\cdot}{\forall} z)x) \leftrightarrow \forall y T(x(\dot{\bar{y}}/z)))]$

T7. $\forall x[Sent(x) \rightarrow (T(\underset{\cdot}{K} x) \leftrightarrow KT(x))]$

The system $\mathbf{EA_T}$ is the theory axiomatized by $\Sigma \cup K\Sigma$, where Σ consists of the axioms of **PA** (in the language of $\mathcal{L}_{EA_T}$), the basic axioms of absolute provability (in the language $\mathcal{L}_{EA_T}$), and the axioms of truth (for the language $\mathcal{L}_{EA_T}$).

Reinhardt showed that in this setting one can prove versions of the incompleteness theorems that pertain to the concept of absolute provability.

Formal First Incompleteness Theorem

Theorem (Reinhardt, 1985)

Assume that **S** include **EA**. Suppose that $F(x)$ is a formula with one free variable and such that for each sentence φ,

$$\mathbf{S} \vdash K(F(\ulcorner \varphi \urcorner) \to \varphi)$$

Then there is a sentence γ such that

$$\mathbf{S} \vdash K\gamma \wedge K\neg F(\ulcorner \gamma \urcorner)$$

W. Reinhardt (1985). *Absolute Versions of Incompleteness Theorems*. Nous, 19(3), pp. 317 - 346.

Formal Second Incompleteness Theorem

Theorem (Reinhardt, 1985)

Assume that **S** include **EA**. Suppose that $F(x)$ is a formula with one free variable and such that for each sentence φ,

$$\mathbf{S} \vdash K(K\varphi \rightarrow F(\ulcorner \varphi \urcorner))$$

Then

$$\mathbf{S} \vdash K \neg K \mathrm{Con}_F$$

W. Reinhardt (1985). *Absolute Versions of Incompleteness Theorems*. Nous, 19(3), pp. 317 - 346.

Formalizing Gödel's Disjunction

- Let '$K_p(x)$' be shorthand for '$T(\underset{\cdot}{K} x) \wedge Sent_{\mathcal{L}_{EA}}(x)$' and
- Let '$x \in W_e$' be shorthand for the statement "x is the eth computably enumerable set."

The First Disjunct:

$$\forall e[\forall x((Sent_{\mathcal{L}_{PA}}(x) \wedge x \in W_e) \rightarrow T(x)) \rightarrow$$
$$\exists y(Sent_{\mathcal{L}_{PA}}(y) \wedge K_p(y) \wedge y \notin W_e)$$

The Second Disjunct:

$$\exists y(Sent_{\mathcal{L}_{PA}}(y) \land T(y) \land \neg T(\underset{\cdot}{K} y) \land \neg T(\underset{\cdot}{K} \underset{\cdot}{\neg} y))$$

Let GD be the disjunction of the previous two sentences.

Theorem (Reinhardt)

1. $\mathbf{EA_T} \vdash \forall x[K_p(x) \rightarrow T(x)]$
2. $\mathbf{EA_T} \vdash \mathrm{GD}$

Proof Sketch

Let $K_p = T$ be shorthand for

$$(\forall x)\,[Sent_{\mathcal{L}_A}(x) \rightarrow (K_p(x) \leftrightarrow T(x))]$$

Proof Sketch

Case 1: $K_p = T$

Because, we allowed 'T' to figure in the induction scheme, the proof of the first incompleteness theorem gives us:

$$(\forall e)[(\forall x)((Sent_{\mathcal{L}_{PA}}(x) \wedge x \in W_e) \rightarrow T(x)) \rightarrow (\exists y)(Sent_{\mathcal{L}_{PA}}(y) \wedge T(y) \wedge y \notin W_e)$$

Since we assumed $K_p = T$, we can replace '$T(y)$' with '$K_p(y)$' to obtain the first disjunct.

Proof Sketch

Case 2: $K_p \neq T$.

$$(\exists x)[Sent_{\mathcal{L}_{PA}}(x) \wedge \neg(K_p(x) \leftrightarrow T(x))]$$

By part 1, it follows that:

$$(\exists x)[Sent_{\mathcal{L}_{PA}}(x) \wedge T(x) \wedge \neg K_p(x)]$$

Fix an instance of x. The last conjunct implies $\neg T(\underset{\cdot}{\mathrm{K}}x)$.

Proof Sketch

But we also have $\neg K_p(\underset{\cdot}{\neg} x)$. (Suppose for contradiction that $K_p(\underset{\cdot}{\neg} x)$ holds. Then $T(\underset{\cdot}{\neg} x)$, which contradicts $T(x)$.)

Since $Sent_{\mathcal{L}_{PA}}(\underset{\cdot}{\neg} x)$ it follows that $\neg T(\underset{\cdot}{\mathsf{K}} \underset{\cdot}{\neg} x)$. Thus, we have shown that

$$(\exists x)[Sent_{\mathcal{L}_{PA}}(x) \wedge T(x) \wedge \neg T(\underset{\cdot}{\mathsf{K}} x) \wedge \neg T(\underset{\cdot}{\mathsf{K}} \underset{\cdot}{\neg} x)]$$

Which Disjunct?

1. $\forall e[\forall x((Sent_{\mathcal{L}_{PA}}(x) \land x \in W_e) \rightarrow T(x)) \rightarrow \exists y(Sent_{\mathcal{L}_{PA}}(y) \land K_p(y) \land y \notin W_e)$

2. $\exists y(Sent_{\mathcal{L}_{PA}}(y) \land T(y) \land \neg T(\underset{\cdot}{K} y) \land \neg T(\underset{\cdot}{K} \underset{\cdot}{\neg} y))$

Which Disjunct?

1. $\forall e[\forall x((Sent_{\mathcal{L}_{PA}}(x) \wedge x \in W_e) \rightarrow T(x)) \rightarrow$
   $\exists y(Sent_{\mathcal{L}_{PA}}(y) \wedge K_p(y) \wedge y \notin W_e)$

2. $\exists y(Sent_{\mathcal{L}_{PA}}(y) \wedge T(y) \wedge \neg T(\underset{\cdot}{K} y) \wedge \neg T(\underset{\cdot}{K}\underset{\cdot}{\neg} y))$

$$\begin{array}{ll}
(\text{WMT}) & \exists e\,(K_p = W_e) \\
(\text{SMT}) & K\exists e\,(K_p = W_e) \\
(\text{SSMT}) & \exists e\,K(K_p = W_e)
\end{array}$$

Which Disjunct?

Theorem (Reinhardt, 1985) $\mathbf{EA_T} + SSMT$ is inconsistent.

William N. Reinhardt (1985). *Absolute versions of incompleteness theorems*. Nous, 19(3), pp. 317-346.

Which Disjunct?

Theorem (Reinhardt, 1985) $\mathbf{EA_T} + \mathit{WMT}$ is consistent.

William N. Reinhardt (1985). *The consistency of a variant of Church's thesis with an axiomatic theory of an epistemic notion*. In Special Volume for the Proceedings of the 5th Latin American Symposium on Mathematical Logic, 1981, vol. 19 of Revista Colombiana de Matem'aticas, pp. 177-200.

Which Disjunct?

Theorem (Carlson, 2005) $\mathbf{EA_T} + SMT$ is consistent.

Timothy J. Carlson (2005). *Knowledge, machines, and the consistency of Reinhardt's strong mechanistic thesis*. Annals of Pure and Applied Logic, 105(1-3), pp. 51-81.

Which Disjunct?

In other words, from the point of view of $\mathbf{EA_T}$ it is entirely possible that the idealized human mind *knows* that it is a Turing machine. It just can't know which one!

A Different Perspective: Type-Free Theory of Truth

Let ‘$D(x)$’ be short for $T(x) \vee T(\underset{\cdot}{\neg} x)$ (which asserts that x is *determinate*)

Axioms of Determinateness

D1. $\forall x\,[Sent_{\mathcal{L}_A}(x) \rightarrow D(x)]$

D2. $\forall x\,[Sent(x) \rightarrow (D(\underset{\cdot}{\neg} x) \leftrightarrow D(x)]$

D3. $\forall x\,\forall y\,[Sent(x) \wedge Sent(y) \rightarrow (D(x \underset{\cdot}{\vee} y) \leftrightarrow (D(x) \wedge D(y)))$

D4. $\forall x\,\forall y\,[Sent(x) \wedge Sent(y) \rightarrow (D(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (D(x) \wedge (T(x) \rightarrow D(y))))$

D5. $\forall x\,\forall z\,[Var(z) \wedge Sent((\underset{\cdot}{\forall} z)x) \rightarrow (D((\underset{\cdot}{\forall} z)x) \leftrightarrow (\forall y\, D(x(\dot{\overline{y}}/z))))]$

D6. $\forall x\,[D(\underset{\cdot}{\mathrm{T}}(\dot{\overline{x}})) \leftrightarrow D(x)]$

Axioms of Truth

T1. For each atomic formula $R(x_1, \ldots, x_n)$:

$$\forall x_1 \cdots \forall x_n [CTerm(x_1) \wedge \cdots \wedge CTerm(x_n) \rightarrow T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \rightarrow R(Val(x_1), \ldots, Val(x_n)]$$

T2. $\forall x[(Sent(x) \wedge D(x)) \rightarrow T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x)]$

T3. $\forall x \forall y[(Sent(x) \wedge Sent(y) \wedge D(x \underset{\cdot}{\vee} y)) \rightarrow (T(x \underset{\cdot}{\vee} y) \leftrightarrow (T(x) \vee T(y)))]$

T4. $\forall x \forall y[(Sent(x) \wedge Sent(y) \wedge D(x \underset{\cdot}{\rightarrow} y)) \rightarrow (T(x \underset{\cdot}{\rightarrow} y) \leftrightarrow (T(x) \rightarrow T(y)))]$

T5.

$$\forall x \forall z[(Var(z) \wedge Sent((\underset{\cdot}{\forall} z)x) \wedge D((\underset{\cdot}{\forall} z)x)) \rightarrow (T((\underset{\cdot}{\forall} z)x) \leftrightarrow \forall y T(x(\dot{\overline{y}}/z)))]$$

T6. $\forall x[D(x) \rightarrow (T(\underset{\cdot}{T}\dot{\overline{x}}) \leftrightarrow T(x))]$

Theorem. (Feferman) For each φ in the language of **DT**,

$$\mathbf{DT} \vdash D(\ulcorner\varphi\urcorner) \rightarrow (T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi)$$

S. Feferman (2008). *Axioms for determinateness and truth*. Review ofSymbolic Logic, 1(2), pp. 204-217.

D7. $\forall x[Sent(x) \rightarrow (D(\underset{\cdot}{K}x) \leftrightarrow D(x))]$

T7. $\forall x[Sent(x) \rightarrow (T(\underset{\cdot}{K}x) \leftrightarrow KT(x))]$

D7. $\forall x[Sent(x) \rightarrow (D(\underset{\cdot}{K}x) \leftrightarrow D(x))]$

T7. $\forall x[Sent(x) \rightarrow (T(\underset{\cdot}{K}x) \leftrightarrow KT(x))]$

(I) For each sentence φ, $K\varphi \rightarrow T(\ulcorner\varphi\urcorner)$

Theorem (Koellner). **EADT** is inconsistent.

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner \varphi \urcorner)$$

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner\varphi\urcorner)$$

Claim 1: $\mathbf{EADT} \vdash K\varphi \rightarrow K_p(\ulcorner\varphi\urcorner)$

1. $K\varphi \rightarrow KK\varphi$ E4

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner\varphi\urcorner)$$

Claim 1: $\mathbf{EADT} \vdash K\varphi \to K_p(\ulcorner\varphi\urcorner)$

1. $K\varphi \to KK\varphi$ E4

2. $KK\varphi \to T(\ulcorner K\varphi\urcorner)$ I

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner \varphi \urcorner)$$

Claim 1: $\mathbf{EADT} \vdash K\varphi \to K_p(\ulcorner \varphi \urcorner)$

1. $K\varphi \to KK\varphi$ E4

2. $KK\varphi \to T(\ulcorner K\varphi \urcorner)$ I

3. $K\varphi \to T(\ulcorner K\varphi \urcorner)$ from 1, 2

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner\varphi\urcorner)$$

Claim 1: $\mathbf{EADT} \vdash K\varphi \rightarrow K_p(\ulcorner\varphi\urcorner)$

1. $K\varphi \rightarrow KK\varphi$ E4

2. $KK\varphi \rightarrow T(\ulcorner K\varphi\urcorner)$ I

3. $K\varphi \rightarrow T(\ulcorner K\varphi\urcorner)$ from 1, 2

4. $K\varphi \rightarrow (Sent_{EADT}(\ulcorner\varphi\urcorner) \wedge T(\ulcorner K\varphi\urcorner))$ φ is a sentence of $\mathcal{L}_{EADT}$

Proof. The main point is that for each φ in the language $\mathcal{L}_{EA_{DT}}$.

$$\mathbf{EADT} \vdash K\varphi \leftrightarrow K_p(\ulcorner \varphi \urcorner)$$

Claim 1: $\mathbf{EADT} \vdash K\varphi \rightarrow K_p(\ulcorner \varphi \urcorner)$

1. $K\varphi \rightarrow KK\varphi$ E4
2. $KK\varphi \rightarrow T(\ulcorner K\varphi \urcorner)$ I
3. $K\varphi \rightarrow T(\ulcorner K\varphi \urcorner)$ from 1, 2
4. $K\varphi \rightarrow (Sent_{EADT}(\ulcorner \varphi \urcorner) \wedge T(\ulcorner K\varphi \urcorner))$ φ is a sentence of $\mathcal{L}_{EADT}$
5. $K\varphi \rightarrow K_p(\ulcorner \varphi \urcorner)$ definition of K_p

Claim 2: $\mathbf{EADT} \vdash K_p(\ulcorner \varphi \urcorner) \to K\varphi$
We want to show that:

$$\mathbf{EADT} \vdash (Sent_{EADT}(\ulcorner \varphi \urcorner) \wedge T(\ulcorner K\varphi \urcorner)) \to K\varphi$$

Claim 2: $\mathbf{EADT} \vdash K_p(\ulcorner \varphi \urcorner) \to K\varphi$
We want to show that:

$$\mathbf{EADT} \vdash (Sent_{EADT}(\ulcorner \varphi \urcorner) \wedge T(\ulcorner K\varphi \urcorner)) \to K\varphi$$

We have that for for all ψ,

$$\mathbf{EADT} \vdash T(\ulcorner \psi \urcorner) \to \psi.$$

Claim 2: $\mathbf{EADT} \vdash K_p(\ulcorner \varphi \urcorner) \to K\varphi$
We want to show that:

$$\mathbf{EADT} \vdash (Sent_{EADT}(\ulcorner \varphi \urcorner) \wedge T(\ulcorner K\varphi \urcorner)) \to K\varphi$$

We have that for for all ψ,

$$\mathbf{EADT} \vdash T(\ulcorner \psi \urcorner) \to \psi.$$

Hence,

$$\mathbf{EADT} \vdash T(\ulcorner K\varphi \urcorner) \to K\varphi.$$

Thus, we can turn the operator K into a predicate K_p and show that K_p has the properties of the Montague-Kaplan Theorem, showing that the theory is inconsistent.

The trouble is that once we enter the realm where indeterminate sentences arise, it is no longer plausible to maintain that one knows logical validities if the validities in question are themselves indeterminate.

The trouble is that once we enter the realm where indeterminate sentences arise, it is no longer plausible to maintain that one knows logical validities if the validities in question are themselves indeterminate.

This motivates a natural modification of the above system. And once one makes this modification it turns out to be possible to treat 'K' as a predicate. On this approach the work in circumventing the paradoxes is carried by the theory of truth and transferred over to the theory of absolute provability.

DTK

The language of **DTK** is $\mathcal{L}_A$ with predicates 'K' (absolute knowledge) and 'T' (truth).

Let $D(x)$ be short for $T(x) \vee T(\underset{\cdot}{\neg} x)$ (determinitness).

The system **DTK** has four groups of axioms (in addition to our fixed set of axioms for first-order logic).

DTK: Arithmetic Axioms

The axioms of arithmetic are those of **PA**, where the induction scheme is extended to the entire language so that 'T' and 'K' are allowed to figure in induction.

DTK: Determinitness Axioms

D1. $\forall x\, [\mathit{At\text{-}Sent}_{\mathcal{L}_A}(x) \to D(x)]$

D2. $\forall x\, [\mathit{Sent}(x) \to (D(\underset{.}{\neg} x) \leftrightarrow D(x)]$

D3. $\forall x\, \forall y\, [\mathit{Sent}(x) \wedge \mathit{Sent}(y) \to (D(x \underset{.}{\vee} y) \leftrightarrow (D(x) \wedge D(y)))$

D4. $\forall x\, \forall y\, [\mathit{Sent}(x) \wedge \mathit{Sent}(y) \to (D(x \underset{.}{\to} y) \leftrightarrow (D(x) \wedge (T(x) \to D(y))))$

D5. $\forall x\, \forall z\, [\mathit{Var}(z) \wedge \mathit{Sent}((\underset{.}{\forall} z)x) \to (D((\underset{.}{\forall} z)x) \leftrightarrow (\forall y\, D(x(\dot{\overline{y}}/z))))]$

D6. $\forall x\, [D(\underset{.}{\mathrm{T}}(\dot{\overline{x}})) \leftrightarrow D(x)]$

D7. $\forall x\, [D(\underset{.}{\mathrm{K}}(\dot{\overline{x}})) \leftrightarrow D(x)]$

DTK: Truth Axioms

T1. For each atomic formula $R(x_1, \ldots, x_n)$:

$$\forall x_1 \cdots \forall x_n [CTerm(x_1) \wedge \cdots \wedge CTerm(x_n) \to T(\underset{\cdot}{R}(x_1, \ldots, x_n)) \to R(Val(x_1), \ldots, Val(x_n))]$$

T2. $\forall x[(Sent(x) \wedge D(x)) \to T(\underset{\cdot}{\neg} x) \leftrightarrow \neg T(x)]$

T3. $\forall x \forall y[(Sent(x) \wedge Sent(y) \wedge D(x \underset{\cdot}{\vee} y)) \to (T(x \underset{\cdot}{\vee} y) \leftrightarrow (T(x) \vee T(y)))]$

T4. $\forall x \forall y[(Sent(x) \wedge Sent(y) \wedge D(x \underset{\cdot}{\to} y)) \to (T(x \underset{\cdot}{\to} y) \leftrightarrow (T(x) \to T(y)))]$

T5. $\forall x \forall z[(Var(z) \wedge Sent((\underset{\cdot}{\forall} z)x) \wedge D((\underset{\cdot}{\forall} z)x)) \to (T((\underset{\cdot}{\forall} z)x) \leftrightarrow \forall y T(x(\dot{\bar{y}}/z)))]$

T6. $\forall x[D(x) \to (T(\underset{\cdot}{T}\dot{\bar{x}}) \leftrightarrow T(x))]$

T7. $\forall x[D(x) \to (T(\underset{\cdot}{K}\dot{\bar{x}}) \leftrightarrow K(x))]$

DTK: Knowledge Axioms

K1. $\forall x\,[Sent(x) \to (K(x) \to T(x))]$

K2. $\forall x\,\forall y\,[(Sent(x) \wedge Sent(y)) \to ((K(x \underset{\cdot}{\to} y) \wedge K(x)) \to K(y))]$

K3. $\forall x\,[Sent(x) \to (K(x) \to K(\underset{\cdot}{K}(\dot{\bar{x}})))]$

DTK: Rules

$$\frac{\varphi \wedge D(\ulcorner \varphi \urcorner)}{K(\ulcorner \varphi \urcorner)}$$

$$\frac{\varphi \wedge D(\ulcorner \varphi \urcorner)}{T(\ulcorner \varphi \urcorner)}$$

Theorem DTK is consistent.

Theorem DTK is consistent.

Proof idea. Feferman used a fixed-point construction to show that **DT** is consistent. We show that **DTK** holds in this model by interpreting **DTK** in **DT**. The interpretation is simply the one that interprets 'K' as 'T'.

Fixed Point Theorem

Theorem. For each $\psi(x)$ in $\mathcal{L}_A$, there exists a sentence φ in $\mathcal{L}_A$ such that

$$\mathbf{DTK} \vdash K(\ulcorner \varphi \leftrightarrow \psi(\ulcorner \varphi \urcorner) \urcorner)$$

First Incompleteness Theorem

Theorem. Suppose that **S** includes **DTK**. Suppose $F(x)$ is a formula in $\mathcal{L}_A$ such that for all sentences φ of $\mathcal{L}_A$:

$$\mathbf{S} \vdash K(\ulcorner F(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$$

Then there is a sentence ψ of $\mathcal{L}_A$ such that

$$\mathbf{S} \vdash K(\ulcorner \psi \urcorner) \wedge K(\ulcorner \neg F(\ulcorner \psi \urcorner) \urcorner)$$

Proof Sketch

1. $\mathbf{S} \vdash K(\ulcorner \psi \leftrightarrow \neg F(\ulcorner \psi \urcorner) \urcorner)$ K-FPT

Proof Sketch

1. $\mathbf{S} \vdash K(\ulcorner \psi \leftrightarrow \neg F(\ulcorner \psi \urcorner) \urcorner)$ K-FPT
2. $\mathbf{S} \vdash K(\ulcorner F(\ulcorner \psi \urcorner) \rightarrow \psi \urcorner)$ Assumption

Proof Sketch

1. $\mathbf{S} \vdash K(\ulcorner \psi \leftrightarrow \neg F(\ulcorner \psi \urcorner) \urcorner)$ *K*-FPT
2. $\mathbf{S} \vdash K(\ulcorner F(\ulcorner \psi \urcorner) \to \psi \urcorner)$ Assumption
3. $\mathbf{S} \vdash K(\ulcorner (F(\ulcorner \psi \urcorner) \to \psi) \to ((\neg F(\ulcorner \psi \urcorner) \to \psi) \to \psi) \urcorner)$ *DK*-Intro

Proof Sketch

1. $\mathbf{S} \vdash K(\ulcorner \psi \leftrightarrow \neg F(\ulcorner \psi \urcorner) \urcorner)$ *K*-FPT

2. $\mathbf{S} \vdash K(\ulcorner F(\ulcorner \psi \urcorner) \rightarrow \psi \urcorner)$ Assumption

3. $\mathbf{S} \vdash K(\ulcorner (F(\ulcorner \psi \urcorner) \rightarrow \psi) \rightarrow ((\neg F(\ulcorner \psi \urcorner) \rightarrow \psi) \rightarrow \psi) \urcorner)$ *DK*-Intro

4. $\mathbf{S} \vdash K(\ulcorner \psi \urcorner)$ Modal Reasoning

Proof Sketch

1.	$\mathbf{S} \vdash K(\ulcorner \psi \leftrightarrow \neg F(\ulcorner \psi \urcorner) \urcorner)$	K-FPT
2.	$\mathbf{S} \vdash K(\ulcorner F(\ulcorner \psi \urcorner) \to \psi \urcorner)$	Assumption
3.	$\mathbf{S} \vdash K(\ulcorner (F(\ulcorner \psi \urcorner) \to \psi) \to ((\neg F(\ulcorner \psi \urcorner) \to \psi) \to \psi) \urcorner)$	DK-Intro
4.	$\mathbf{S} \vdash K(\ulcorner \psi \urcorner)$	Modal Reasoning
5.	$\mathbf{S} \vdash K(\ulcorner \neg F(\ulcorner \psi \urcorner) \urcorner)$	Modal Reasoning

Second Incompleteness Theorem

Theorem. Suppose that **S** includes **DTK**. Suppose $F(x)$ is a formula in $\mathcal{L}_A$ such that for all sentences φ of $\mathcal{L}_A$:

$$\mathbf{S} \vdash K(\ulcorner K(\ulcorner \varphi \urcorner) \to F(\ulcorner \varphi \urcorner) \urcorner)$$

Then,

$$\mathbf{S} \vdash K(\ulcorner \neg K(\ulcorner Con(F) \urcorner) \urcorner)$$

Theorem. **DTK** can prove the formalized version of Gödel's disjunction.

Theorem. **DTK** can prove the formalized version of Gödel's disjunction.

Theorem. Assume that **DTK** is consistent. Then **DTK** can neither refute the first disjunct nor prove the second disjunct of Gödel's Disjunction.

Theorem. Assume that **DTK** is Σ_n sound for all $n < \omega$. Then **DTK** can neither prove the first disjunct nor refute the second disjunct of Gödel's Disjunction.

The above consistency results show that the principles of **DTK** are insufficient to prove or refute either disjunct; indeed they are insufficient to distinguish between the "*K* equals truth" interpretation (rational optimism) and the "*K* equals relative provability" interpretation (pessimistic mechanism).

One natural response to this—on behalf of the proponent of the first disjunct—is that the principles of knowledge embodied in **DTK** are merely a fragment of the acceptable principles of knowledge, and that when we supplement **DTK** with the missing principles of knowledge we will be in a position to prove or refute the disjuncts.

...it seems likely that the above independence results will carry over to new frameworks when one incorporates new theories of knowledge and type-free truth. And for this reason it seems likely that both the question of whether "the mind can be mechanized" and the question of whether "there are absolutely undecidable statements" (in the idealized sense we have been considering) are themselves examples of "absolutely undecidable statements" and, as such, will remain forever undecided and continue to lie outside the scope of human reason. *(Koellner, p. 184-5)*

More on Gödel's Disjunction

P. Koellner (2018). *On the Question of Whether the Mind can be Mechanized I: From Gödel to Penrose*. Journal of Philosophy, Volume CXV, No. 7, pp. 337 - 360.

P. Koellner (2010). *On the Question of Absolute Undecidability*. in *Kurt Gödel: Essays for his Centennial*, Solomon Feferman, Charles Parsons, and Stephen G. Simpson (eds.), Lecture Notes in Logic, 33.

Wesley Wrigley (2022). *Gödel's Disjunctive Argument*. Philosophia Mathematica, pp. 306-342.

More on Epistemic Arithmetic

S. Shapiro (1985). *Epistemic and intuitionistic arithmetic*. in *Intensional Mathematics*, S. Shapiro (ed.), North-Holland.

L. Horsten (1994). *Modal-Epistemic Variants of Shapiro's System of Epistemic Arithmetic*. Notre Dame Journal of Formal Logic, 35(2).

L. Horsten (1998). *In Defense of Epistemic Arithmetic*. Synthese, 116, pp. 1 - 25.

Epistemic Arithmetic and Heyting Arithmetic

S. Shapiro (1985). *Epistemic and intuitionistic arithmetic*. in *Intensional Mathematics*, S. Shapiro (ed.), North-Holland.

Nicolas D. Goodman (1984). *Epistemic Arithmetic is a Conservative Extension of Intuitionistic Arithmetic*. The Journal of Symbolic Logic, 49(1), pp. 192 - 203.

R. C. Flagg and H. Friedman (1986). *Epistemic and Inuitionistic Formal Systems*. Annals of Pure and Applied Logic, 32, pp. 53-60.

Plan

- ✓ Introduction: Smullyan's Machine
- ✓ Background
  - ✓ Formal Arithmetic
  - ✓ Gödel's Incompleteness Theorems
  - ✓ Names and Gödel numbering
  - ✓ Fixed Point Theorem
- ✓ Provability predicate and Löb's Theorem
- ✓ Provability logic
- ✓ Predicate approach to modality
- ✓ The Knower Paradox and variants
- ✓ A Primer on Epistemic and Doxastic Logic
- ✓ Anti-Expert Paradox, and related paradoxes
- ✓ Predicate approach to modality, continued
- ✓ Epistemic Arithmetic
- ✓ Gödel's Disjunction

Thank you!

https://pacuit.org/esslli2025/epistemic-arithmetic/