Epistemic Arithmetic

Eric Pacuit
University of Maryland

Lecture 2, ESSLLI 2025

July 29, 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
- ▶ A Primer on Epistemic and Doxastic Logic
- ▶ Anti-Expert Paradoxes
- ▶ The Knower Paradox and variants
- ▶ Epistemic Arithmetic
- ▶ Gödel's Disjunction

H. Gaifman (2006). *Naming and Diagonalization, From Cantor to Gödel to Kleene*. Logic Journal of the IGPL, pp. 709 - 728.

What's in a name?

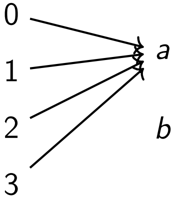
Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

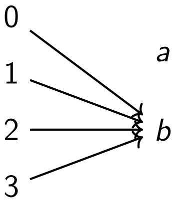
	0	1	2	3
	a	a	a	a



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

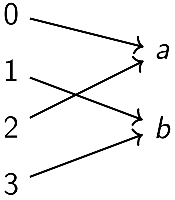
	0	1	2	3
	a	a	a	a
	b	b	b	b



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

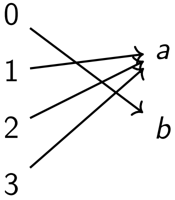
	0	1	2	3
	a	a	a	a
	b	b	b	b
	a	b	a	b



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

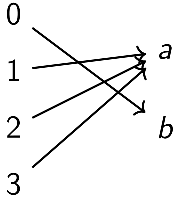
	0	1	2	3
	a	a	a	a
	b	b	b	b
	a	b	a	b
	b	a	a	a



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$

0 $\qquad$ a

1

2 $\qquad$ b

3

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$

0

1

2

3

a

b

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$

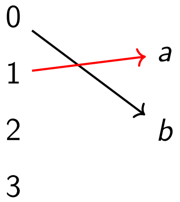
$$\begin{tikzcd} 0 \arrow[dddr] & \\ & a \\ 1 & \\ 2 & b \\ 3 & \end{tikzcd}$$

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$

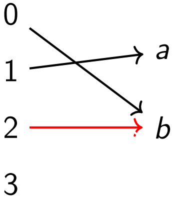


What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$

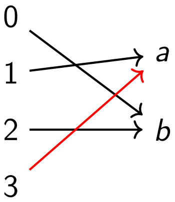


What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
α	a	a	a	a
β	b	b	b	b
γ	a	b	a	b
δ	b	a	a	a

$$g(n) = \begin{cases} b & \text{if } \gamma(n) = a \\ a & \text{if } \gamma(n) = b \end{cases}$$



What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

0	
1	a
2	b
3	

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

$$diag(n) = \begin{cases} b & \text{if } [n](n) = a \\ a & \text{if } [n](n) = b \end{cases}$$

0

a

1

2

b

3

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

$$diag(n) = \begin{cases} b & \text{if } [n](n) = a \\ a & \text{if } [n](n) = b \end{cases}$$

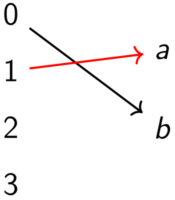
$$\begin{array}{ccc} 0 & \longrightarrow & b \\ 1 & & a \\ 2 & & \\ 3 & & \end{array}$$

What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

$$diag(n) = \begin{cases} b & \text{if } [n](n) = a \\ a & \text{if } [n](n) = b \end{cases}$$

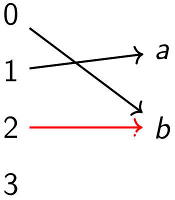


What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

$$diag(n) = \begin{cases} b & \text{if } [n](n) = a \\ a & \text{if } [n](n) = b \end{cases}$$

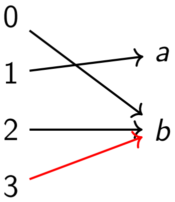


What's in a name?

Functions from $\{0, 1, 2, 3\}$ to $\{a, b\}$

	0	1	2	3
[0]	a	a	a	a
[1]	b	b	b	b
[2]	a	b	a	b
[3]	b	a	a	a

$$diag(n) = \begin{cases} b & \text{if } [n](n) = a \\ a & \text{if } [n](n) = b \end{cases}$$



Cantor's Diagonalization Proof

Functions from $\mathbb{N}$ to $\{0, 1\}$

	0	1	2	3	$\cdots$	n	$\cdots$
	0	0	0	0	$\cdots$	0	$\cdots$
	0	1	0	1	$\cdots$	1	$\cdots$
	0	1	1	0	$\cdots$	0	$\cdots$
	1	0	1	0	$\cdots$	1	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
	0	0	1	0	$\cdots$	1	$\cdots$
	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Cantor's Diagonalization Proof

Functions from $\mathbb{N}$ to $\{0,1\}$

$$
\begin{array}{c|cccccccc}
 & 0 & 1 & 2 & 3 & \cdots & n & \cdots \\
\hline
[0] & \boxed{0} & 0 & 0 & 0 & \cdots & 0 & \cdots \\
[1] & 0 & \boxed{1} & 0 & 1 & \cdots & 1 & \cdots \\
[2] & 0 & 1 & \boxed{1} & 0 & \cdots & 0 & \cdots \\
[3] & 1 & 0 & 1 & \boxed{0} & \cdots & 1 & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
[n] & 0 & 0 & 1 & 0 & \cdots & \boxed{1} & \cdots \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots
\end{array}
$$

$d : \mathbb{N} \to \{0,1\}$ $\qquad d(n) = 1 - [n](n)$

Cantor's Diagonalization Proof

Functions from $\mathbb{N}$ to $\{0,1\}$

	0	1	2	3	$\cdots$	n	$\cdots$
$[0]$	0	0	0	0	$\cdots$	0	$\cdots$
$[1]$	0	1	0	1	$\cdots$	1	$\cdots$
$[2]$	0	1	1	0	$\cdots$	0	$\cdots$
$[3]$	1	0	1	0	$\cdots$	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$[n]$	0	0	1	0	$\cdots$	1	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$d : \mathbb{N} \to \{0,1\} \qquad d(n) = 1 - [n](n)$$

Then, $d \neq [n]$ for any $n \in \mathbb{N}$.

Cantor's original statement is phrased as a non-existence claim: there is no function mapping all the members of a set S onto the set of all $0, 1$-valued functions over S. But the proof establishes a positive result: given any way of correlating functions with members of S, one can construct a function not correlated with any member of S.

(Gaiffman, p. 711)

Richard's Paradox (1905)

Consider all the definitions (in English) of real numbers.

Richard's Paradox (1905)

Consider all the definitions (in English) of real numbers. Since any such definition is a finite sequence of letters, the definitions can be listed in order.

Richard's Paradox (1905)

Consider all the definitions (in English) of real numbers. Since any such definition is a finite sequence of letters, the definitions can be listed in order.

Let u_i be the real number defined by the ith definition and $f_i(n)$ be the nth member of the decimal expansion of u_i.

Richard's Paradox (1905)

Consider all the definitions (in English) of real numbers. Since any such definition is a finite sequence of letters, the definitions can be listed in order.

Let u_i be the real number defined by the ith definition and $f_i(n)$ be the nth member of the decimal expansion of u_i.

Let u^* be the number who's decimal expansion is $0.g(1)g(2)\cdots g(n)\cdots$ where g is defined by $g(n) = f_n(n) + 1$ if $f_n(n) < 8$, $g(n) = 1$ otherwise.

Richard's Paradox (1905)

Consider all the definitions (in English) of real numbers. Since any such definition is a finite sequence of letters, the definitions can be listed in order.

Let u_i be the real number defined by the ith definition and $f_i(n)$ be the nth member of the decimal expansion of u_i.

Let u^* be the number who's decimal expansion is $0.g(1)g(2)\cdots g(n)\cdots$ where g is defined by $g(n) = f_n(n) + 1$ if $f_n(n) < 8$, $g(n) = 1$ otherwise.

But the previous description defines a number, so $u^* = u_i$ for some i. But, this is impossible.

Richard's Paradox (1905)

1. Let A be the set of all positive integers that can be defined in under 100 words. Since there are only finitely many of these, there must be a smallest positive integer n that does not belong to A.

Richard's Paradox (1905)

1. Let A be the set of all positive integers that can be defined in under 100 words. Since there are only finitely many of these, there must be a smallest positive integer n that does not belong to A.

   But haven't I just defined n in under 100 words?

Richard's Paradox (1905)

1. Let A be the set of all positive integers that can be defined in under 100 words. Since there are only finitely many of these, there must be a smallest positive integer n that does not belong to A.

   But haven't I just defined n in under 100 words?

2. Let B be the set of all reasonably interesting positive integers. Let n be the smallest integer not belonging to B.

Richard's Paradox (1905)

1. Let A be the set of all positive integers that can be defined in under 100 words. Since there are only finitely many of these, there must be a smallest positive integer n that does not belong to A.

   But haven't I just defined n in under 100 words?

2. Let B be the set of all reasonably interesting positive integers. Let n be the smallest integer not belonging to B.

   But surely this defining property of n makes it reasonably interesting.

Let f be a function that associates each number $x \in \mathbb{N}$ with a subset of $\mathbb{N}$, i.e., for all $x \in \mathbb{N}$, $f(x) \subseteq \mathbb{N}$.

Let f be a function that associates each number $x \in \mathbb{N}$ with a subset of $\mathbb{N}$, i.e., for all $x \in \mathbb{N}$, $f(x) \subseteq \mathbb{N}$.

Define S^* by:

$$x \in S^* \Leftrightarrow x \notin f(x)$$

Let f be a function that associates each number $x \in \mathbb{N}$ with a subset of $\mathbb{N}$, i.e., for all $x \in \mathbb{N}$, $f(x) \subseteq \mathbb{N}$.

Define S^* by:

$$x \in S^* \Leftrightarrow x \notin f(x)$$

The assumption that there is some z such that $f(z) = S^*$ leads to a contradiction.

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	0	0	0	0	$\cdots$	0	$\cdots$	
$f(1)$	0	1	0	1	$\cdots$	1	$\cdots$	
$f(2)$	0	1	1	0	$\cdots$	0	$\cdots$	
$f(3)$	1	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(n)$	0	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$f(1)$	0	1	0	1	$\cdots$	1	$\cdots$	
$f(2)$	0	1	1	0	$\cdots$	0	$\cdots$	
$f(3)$	1	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(n)$	0	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$f(1)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$f(2)$	0	1	1	0	$\cdots$	0	$\cdots$	
$f(3)$	1	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(n)$	0	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$f(1)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$f(2)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$f(3)$	1	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(n)$	0	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$f(1)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$f(2)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$f(3)$	1	0	1	0	$\cdots$	1	$\cdots$	$\{0, 2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	
$f(n)$	0	0	1	0	$\cdots$	1	$\cdots$	
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$f(0)$	**0**	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$f(1)$	0	**1**	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$f(2)$	0	1	**1**	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$f(3)$	1	0	1	**0**	$\cdots$	1	$\cdots$	$\{0, 2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$f(n)$	0	0	1	0	$\cdots$	**1**	$\cdots$	$\{2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$n \in S^* \text{ iff } n \notin f(n)$$

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$\varphi_0(x)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$\varphi_1(x)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$\varphi_2(x)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$\varphi_3(x)$	1	0	1	0	$\cdots$	1	$\cdots$	$\{0, 2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\varphi_n(x)$	0	0	1	0	$\cdots$	1	$\cdots$	$\{2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$n \in S^*$ iff $n \notin$ set defined by $\varphi_n(x)$

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$\varphi_0(x)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$\varphi_1(x)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \dots, n, \dots\}$
$\varphi_2(x)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$\varphi_3(x)$	1	0	1	0	$\cdots$	1	$\cdots$	$\{0, 2, \dots, n, \dots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\varphi_n(x)$	0	0	1	0	$\cdots$	1	$\cdots$	$\{2, \dots, n, \dots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$$n \in S^* \text{ iff } n \notin \text{ set defined by } \varphi_n(x)$$

Suppose that S^* is definable in our language (say by $\varphi_m(x)$)

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$\varphi_0(x)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$\varphi_1(x)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$\varphi_2(x)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$\varphi_3(x)$	1	0	1	0	$\cdots$	1	$\cdots$	$\{0, 2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\varphi_n(x)$	0	0	1	0	$\cdots$	1	$\cdots$	$\{2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$n \in S^*$ iff $n \notin$ set defined by $\varphi_n(x)$

Write $\varphi_m(\overline{\mathrm{n}})$ for "$\varphi_m(x)$ is true of n"

	0	1	2	3	$\cdots$	n	$\cdots$	$S \subseteq \mathbb{N}$
$\varphi_0(x)$	0	0	0	0	$\cdots$	0	$\cdots$	$\emptyset$
$\varphi_1(x)$	0	1	0	1	$\cdots$	1	$\cdots$	$\{1, 3, \ldots, n, \ldots\}$
$\varphi_2(x)$	0	1	1	0	$\cdots$	0	$\cdots$	$\{1, 2\}$
$\varphi_3(x)$	1	0	1	0	$\cdots$	1	$\cdots$	$\{0, 2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\varphi_n(x)$	0	0	1	0	$\cdots$	1	$\cdots$	$\{2, \ldots, n, \ldots\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

$n \in S^*$ iff $n \notin$ set defined by $\varphi_n(x)$

$$\varphi_m(\overline{n}) \leftrightarrow \neg\mathrm{True}(\ulcorner \varphi_n(\overline{n}) \urcorner)$$

where $\ulcorner \varphi_n(\overline{n}) \urcorner$ is the term in the language representing the code of $\varphi_n(\overline{n})$

D-Liar

$$\varphi_m(\overline{\mathrm{m}}) \leftrightarrow \neg\mathrm{True}(\ulcorner \varphi_m(\overline{\mathrm{m}}) \urcorner)$$

"m is true of $\varphi_m(x)$ iff it is not true that m is true of $\varphi_m(x)$"

Gödel's Idea

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg\mathsf{True}(\ulcorner \varphi_m(\overline{\mathsf{m}}) \urcorner)$$

Gödel's Idea

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg\mathsf{True}(\ulcorner\varphi_m(\overline{\mathsf{m}})\urcorner)$$

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg\mathsf{Prov}(\ulcorner\varphi_m(\overline{\mathsf{m}})\urcorner)$$

"$\varphi_m(\overline{\mathsf{m}})$ is true iff $\varphi_m(\overline{\mathsf{m}})$ is not provable."

Gödel's Idea

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg \mathsf{True}(\ulcorner \varphi_m(\overline{\mathsf{m}}) \urcorner)$$

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg \mathsf{Prov}(\ulcorner \varphi_m(\overline{\mathsf{m}}) \urcorner)$$

"$\varphi_m(\overline{\mathsf{m}})$ is true iff $\varphi_m(\overline{\mathsf{m}})$ is not provable."

$$\varphi_m(\overline{\mathsf{m}}) \leftrightarrow \neg\mathsf{Prov}(\ulcorner \varphi_m(\overline{\mathsf{m}}) \urcorner)$$

$$\varphi_m(\overline{\mathrm{m}}) \leftrightarrow \neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$$

$\varphi_m(\overline{\mathrm{m}})$ is not provable: Suppose $\varphi_m(\overline{\mathrm{m}})$ is provable. Then, since we can only prove true statements, $\varphi_m(\overline{\mathrm{m}})$ is true. This means that $\neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$ is true. So, $\varphi_m(\overline{\mathrm{m}})$ is not provable. Contradiction.

$\varphi_m(\overline{\mathrm{m}}) \leftrightarrow \neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$

$\varphi_m(\overline{\mathrm{m}})$ is not provable: Suppose $\varphi_m(\overline{\mathrm{m}})$ is provable. Then, since we can only prove true statements, $\varphi_m(\overline{\mathrm{m}})$ is true. This means that $\neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$ is true. So, $\varphi_m(\overline{\mathrm{m}})$ is not provable. Contradiction.

$\neg\varphi_m(\overline{\mathrm{m}})$ is not provable: Suppose that $\neg\varphi_m(\overline{\mathrm{m}})$ is provable. Since our system only proves true statements, $\neg\varphi_m(\overline{\mathrm{m}})$ is true. Then $\neg\neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$ is true. So, $\varphi_m(\overline{\mathrm{m}})$ is provable. This contradicts the assumption that the system is consistent.

$\varphi_m(\overline{\mathrm{m}}) \leftrightarrow \neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$

$\varphi_m(\overline{\mathrm{m}})$ is not provable: Suppose $\varphi_m(\overline{\mathrm{m}})$ is provable. Then, since we can only prove true statements, $\varphi_m(\overline{\mathrm{m}})$ is true. This means that $\neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$ is true. So, $\varphi_m(\overline{\mathrm{m}})$ is not provable. Contradiction.

$\neg\varphi_m(\overline{\mathrm{m}})$ is not provable: Suppose that $\neg\varphi_m(\overline{\mathrm{m}})$ is provable. Since our system only proves true statements, $\neg\varphi_m(\overline{\mathrm{m}})$ is true. Then $\neg\neg\mathrm{Prov}(\ulcorner\varphi_m(\overline{\mathrm{m}})\urcorner)$ is true. So, $\varphi_m(\overline{\mathrm{m}})$ is provable. This contradicts the assumption that the system is consistent.

Conclusion: Neither $\varphi_m(\overline{\mathrm{m}})$ nor $\neg\varphi_m(\overline{\mathrm{m}})$ is provable.

$$\varphi_m(\overline{\mathrm{m}}) \leftrightarrow \neg \mathrm{Prov}(\ulcorner \varphi_m(\overline{\mathrm{m}}) \urcorner)$$

1. Apply Richard's move to Cantor's construction to get the D-Liar
2. Replace 'true' with 'provable' on the right-hand side of the sentence
3. Proceed with the difficult task of *arithmetizing syntax* to construct the right-side of the sentence ($\mathrm{Prov}(v)$).
4. Show that the above sentence is provable within the formal system eliminating any appeal to the concept of "truth". The assumption that provable implies truth is replaced with (ω-)consistency.

H. Gaifman (2006). *Naming and Diagonalization, From Cantor to Gödel to Kleene*. Logic Journal of the IGPL, pp. 709 - 728.

Naming systems

Naming systems are intended as a basic framework for studying situations in which functions can be applied to their names....In a naming system we do not specify how the names are attached to functions, we assume only that there is such a correlation and that it satisfies certain minimal requirements.

H. Gaifman (2006). *Naming and Diagonalization, From Cantor to Gödel to Kleene*. Logic Journal of the IGPL, pp. 709 - 728.

Naming systems I

$$\mathcal{D} = (D, \mathit{type}, \{\ \})$$

such that:

- D is a non-empty set.
- type assigns to each $a \in D$ its type: $\mathit{type}(a)$ tells us if a is a name (of a function) and, if it is, the function's arity.

  A name of arity n, or n-ary name, is one that names an n-ary function.

  Types can be construed as tuples: (0)—if a is not a name, $(1, n)$—if it is an n-ary name.
- $\{\ \}$ is a mapping that assigns to every n-ary name, a, a function:

$$\{a\} : D^n \to D$$

Naming systems II

- There is at least one named function of arity greater than 0

Naming systems II

- There is at least one named function of arity greater than 0
- Substitution of names (SN): If f is an n-ary named function, where $n > 0$, then, for every name a:

$$\lambda x_2, \ldots x_n f(a, x_2, \ldots, x_n) \text{ is named}$$

Naming systems II

- There is at least one named function of arity greater than 0

- Substitution of names (SN): If f is an n-ary named function, where $n > 0$, then, for every name a:

$$\lambda x_2, \ldots x_n f(a, x_2, \ldots, x_n) \text{ is named}$$

- Variable permutation (VP): If f is an n-ary named function, where $n > 0$, and π is a permutation of $\{1, \ldots, n\}$, then

$$\lambda x_1, \ldots x_n f(x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(n)}) \text{ is named}$$

n-Diagonal Function

For $n > 0$, an n-diagonal function, denoted dl_n, is a function that maps each n-ary name a to a name of the function:

$$\lambda x_2, \ldots, x_n\{a\}(a, x_2, \ldots, x_n)$$

Thus, $dl_n(a)$ is the name of the above function.

n-Diagonal Function

For $n > 0$, an n-diagonal function, denoted dl_n, is a function that maps each n-ary name a to a name of the function:

$$\lambda x_2, \ldots, x_n\{a\}(a, x_2, \ldots, x_n)$$

Thus, $dl_n(a)$ is the name of the above function.

For all n-ary names a,

$$\{dl_n(a)\}(x_2, \ldots, x_n) = \{a\}(a, x_2, \ldots, x_n)$$

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dI_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

Proof. Let c be the name of $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ and let $e = dl_{n+1}(c)$. Then for and $\vec{x} = (x_1, \ldots, x_n)$,

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dI_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

Proof. Let c be the name of $F(dI_{n+1}(x_0), x_1, \ldots, x_n)$ and let $e = dI_{n+1}(c)$. Then for and $\vec{x} = (x_1, \ldots, x_n)$,

$$\{e\}(\vec{x}) \quad = \quad \{dI_{n+1}(c)\}(\vec{x}) \quad \text{(definition of } e\text{)}$$

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

Proof. Let c be the name of $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ and let $e = dl_{n+1}(c)$. Then for and $\vec{x} = (x_1, \ldots, x_n)$,

$$\begin{aligned} \{e\}(\vec{x}) &= \{dl_{n+1}(c)\}(\vec{x}) && \text{(definition of } e\text{)} \\ &= \{c\}(c, \vec{x}) && \text{(definition of } dl_{n+1}(c)\text{)} \end{aligned}$$

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

Proof. Let c be the name of $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ and let $e = dl_{n+1}(c)$. Then for and $\vec{x} = (x_1, \ldots, x_n)$,

$$\begin{array}{rcll} \{e\}(\vec{x}) & = & \{dl_{n+1}(c)\}(\vec{x}) & \text{(definition of } e\text{)} \\ & = & \{c\}(c, \vec{x}) & \text{(definition of } dl_{n+1}(c)\text{)} \\ & = & F(dl_{n+1}(c), \vec{x}) & \text{(definition of } c\text{)} \end{array}$$

General Fixed-Point Theorem

GFP Theorem. If F is an $(n+1)$-ary named function, $n \geq 0$, and the composition $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ is named, then there is an n-ary name, e, such that:

$$\{e\}(x_1, \ldots, x_n) = F(e, x_1, \ldots, x_n)$$

Proof. Let c be the name of $F(dl_{n+1}(x_0), x_1, \ldots, x_n)$ and let $e = dl_{n+1}(c)$. Then for and $\vec{x} = (x_1, \ldots, x_n)$,

$$\begin{aligned}
\{e\}(\vec{x}) &= \{dl_{n+1}(c)\}(\vec{x}) && \text{(definition of } e\text{)} \\
&= \{c\}(c, \vec{x}) && \text{(definition of } dl_{n+1}(c)\text{)} \\
&= F(dl_{n+1}(c), \vec{x}) && \text{(definition of } c\text{)} \\
&= F(e, \vec{x}) && \text{(definition of } e\text{)}
\end{aligned}$$

- ✓ Gödel numbering
- ✓ Gödel-Carnap Fixed Point Theorem
- ✓ (Naming systems)
- ▶ Representing functions/relations

Representability

Definition

Suppose that $f : \mathbb{N}^k \to \mathbb{N}$. We say that f is **representable** in **Q** when there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$: if $f(n_0, \ldots, n_{k-1}) = m$ then

1. $\mathbf{Q} \vdash A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$
2. $\mathbf{Q} \vdash \forall y(A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, y) \to y = \overline{m})$

Equivalent definitions of representability

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \dots, x_{k-1}, y)$ such that for all $n_0, \dots, n_{k-1} \in \mathbb{N}$, if $f(n_0, \dots, n_{k-1}) = m$ then:

$$\mathbf{Q} \vdash \forall y(A_f(\overline{n_0}, \dots, \overline{n_{k-1}}, y) \leftrightarrow y = \overline{m})$$

Equivalent definitions of representability

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$, if $f(n_0, \ldots, n_{k-1}) = m$ then:

$$\mathbf{Q} \vdash \forall y(A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, y) \leftrightarrow y = \overline{m})$$

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$:
  1. If $f(n_0, \ldots, n_{k-1}) = m$, then $\mathbf{Q} \vdash A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$
  2. If $f(n_0, \ldots, n_{k-1}) \neq m$, then $\mathbf{Q} \vdash \neg A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$

Equivalent definitions of representability

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$, if $f(n_0, \ldots, n_{k-1}) = m$ then:

$$\mathbf{Q} \vdash \forall y(A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, y) \leftrightarrow y = \overline{m})$$

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$:
  1. If $f(n_0, \ldots, n_{k-1}) = m$, then $\mathbf{Q} \vdash A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$
  2. If $f(n_0, \ldots, n_{k-1}) \neq m$, then $\mathbf{Q} \vdash \neg A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$

- f is representable in $\mathbf{Q}$ iff there is a formula $A_f(x_0, \ldots, x_{k-1}, y)$ such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$:
  1. if $f(n_0, \ldots, n_{k-1}) = m$ then $\mathbf{Q} \vdash A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, \overline{m})$
  2. $\mathbf{Q} \vdash \exists! y A_f(\overline{n_0}, \ldots, \overline{n_{k-1}}, y)$

Exercise

Prove that all of the definitions of representability are equivalent.

Representing Relations

A relation $R \subseteq \mathbb{N}^k$ is **representable** in **Q** provided that the characteristic function χ_R is representable in **Q**. It is not hard to see that this is equivalent to saying that $R \subseteq \mathbb{N}^k$ is representable in **Q** provided that there is a formula A_R such that for all $n_0, \ldots, n_{k-1} \in \mathbb{N}$:

1. if $(n_0, \ldots, n_{k-1}) \in R$, then $\mathbf{Q} \vdash A_R(\overline{n_0}, \ldots, \overline{n_{k-1}})$
2. if $(n_0, \ldots, n_{k-1}) \notin R$, then $\mathbf{Q} \vdash \neg A_R(\overline{n_0}, \ldots, \overline{n_{k-1}})$

All of the following relations are representable in **Q**:

- $\mathit{Sent}(x)$: x is the Gödel number of a sentence of $\mathcal{L}_A$
- $\mathit{Form}(x)$: x is the Gödel number of a formula of $\mathcal{L}_A$
- $\mathit{Term}(x)$: x is the Gödel number of a term of $\mathcal{L}_A$
- $\mathit{Axiom}(x)$: x is the Gödel number of an axiom of **Q**
- $\mathit{Prf}_{\mathbf{PA}}(x, y)$: x is the Gödel number of a derivation in **PA** of a formula with Gödel number y.
- $\ldots$

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
- A Primer on Epistemic and Doxastic Logic
- Anti-Expert Paradoxes
- The Knower Paradox and variants
- Epistemic Arithmetic
- Gödel's Disjunction

Proof Predicate

The proof relation $Prf_{\mathbf{PA}}(x, y)$ is represented by a formula $\mathrm{Prf}_{\mathbf{PA}}$.

Proof Predicate

The proof relation $\mathit{Prf}_{\mathbf{PA}}(x, y)$ is represented by a formula $\mathrm{Prf}_{\mathbf{PA}}$.

The *proof predicate*, denoted $\mathrm{Prov}_{\mathbf{PA}}(y)$, is defined as follows:

$$\exists x \mathrm{Prf}_{\mathbf{PA}}(x, y)$$

Derivability Conditions

It can be shown that the provability predicate $\mathrm{Prov}_{\mathbf{PA}}$ satisfies the following:

*D*1. If $\mathbf{PA} \vdash A$, then $\mathbf{PA} \vdash \mathrm{Prov}_{\mathbf{PA}}(\ulcorner A \urcorner)$

*D*2. $\mathbf{PA} \vdash \mathrm{Prov}_{\mathbf{PA}}(\ulcorner A \to B \urcorner) \to (\mathrm{Prov}_{\mathbf{PA}}(\ulcorner A \urcorner) \to \mathrm{Prov}_{\mathbf{PA}}(\ulcorner B \urcorner))$

*D*3. $\mathbf{PA} \vdash \mathrm{Prov}_{\mathbf{PA}}(\ulcorner A \urcorner) \to \mathrm{Prov}_{\mathbf{PA}}(\ulcorner \mathrm{Prov}_{\mathbf{PA}}(\ulcorner A \urcorner) \urcorner)$

Derivability Conditions

A provability predicate for $\mathbf{T}$, denoted $\mathsf{Prov}_{\mathbf{T}}$, satisfies the following:

*D*1. If $\mathbf{T} \vdash A$, then $\mathbf{T} \vdash \mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner)$

*D*2. $\mathbf{T} \vdash \mathsf{Prov}_{\mathbf{T}}(\ulcorner A \to B \urcorner) \to (\mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to \mathsf{Prov}_{\mathbf{T}}(\ulcorner B \urcorner))$

*D*3. $\mathbf{T} \vdash \mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to \mathsf{Prov}_{\mathbf{T}}(\ulcorner \mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \urcorner)$

Reflection Principle

The reflection principle for **T** is the schema

$$\mathsf{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$$

Monotonicity Inference for the Provability Predicate

Lemma

For any theory $\mathbf{T}$, if $\mathrm{Prov}_{\mathbf{T}}$ satisfies $D1$ and $D2$, then:

$$\text{From } \mathbf{T} \vdash A \to B, \text{ infer } \mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to \mathrm{Prov}(\ulcorner B \urcorner).$$

Löb's Theorem

Theorem (Löb's Theorem)

Let $\mathbf{T}$ be an axiomatizable theory extending $\mathbf{Q}$, and suppose $\mathrm{Prov}_{\mathbf{T}}(y)$ is a formula satisfying conditions $D1$-$D3$.

$$\text{If } \mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A, \text{ then } \mathbf{T} \vdash A.$$

Suppose A is a sentence such that $\mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$. Let $B(y)$ be the formula

$$\mathrm{Prov}_{\mathbf{T}}(y) \to A$$

Suppose A is a sentence such that $\mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$. Let $B(y)$ be the formula

$$\mathrm{Prov}_{\mathbf{T}}(y) \to A$$

By the Fixed-Point Theorem, there is a sentence D such that

$$\mathbf{T} \vdash D \leftrightarrow B(\ulcorner D \urcorner)$$

Suppose that $\mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$.

Suppose A is a sentence such that $\mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$. Let $B(y)$ be the formula

$$\mathrm{Prov}_{\mathbf{T}}(y) \to A$$

By the Fixed-Point Theorem, there is a sentence D such that

$$\mathbf{T} \vdash D \leftrightarrow B(\ulcorner D \urcorner)$$

Suppose that $\mathbf{T} \vdash \mathrm{Prov}_{\mathbf{T}}(\ulcorner A \urcorner) \to A$.

To simplify the notation, we write $\mathrm{Prov}(\cdot)$ instead of $\mathrm{Prov}_{\mathbf{T}}$

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \to A)$	FPT
2.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \to A \urcorner)$	Lemma: 1
3.	$\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \to A \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	D2
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner)$	D3

1.	$D \leftrightarrow (\mathsf{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathsf{Prov}(\ulcorner D \urcorner) \to (\mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner)$	PC: 4, 5

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \rightarrow A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \rightarrow (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \rightarrow \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathrm{Prov}(\ulcorner D \urcorner) \rightarrow \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathrm{Prov}(\ulcorner D \urcorner) \rightarrow \mathrm{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathrm{Prov}(\ulcorner A \urcorner) \rightarrow A$	Assumption

1.	$D \leftrightarrow (\mathsf{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathsf{Prov}(\ulcorner D \urcorner) \to (\mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathsf{Prov}(\ulcorner A \urcorner) \to A$	Assumption
8.	$\mathsf{Prov}(\ulcorner D \urcorner) \to A$	PC: 6, 7

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathrm{Prov}(\ulcorner A \urcorner) \to A$	Assumption
8.	$\mathrm{Prov}(\ulcorner D \urcorner) \to A$	PC: 6, 7
9.	D	PC: 1, 8

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathrm{Prov}(\ulcorner A \urcorner) \to A$	Assumption
8.	$\mathrm{Prov}(\ulcorner D \urcorner) \to A$	PC: 6, 7
9.	D	PC: 1, 8
10.	$\mathrm{Prov}(\ulcorner D \urcorner)$	D1 from 9

1.	$D \leftrightarrow (\mathrm{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathrm{Prov}(\ulcorner D \urcorner) \to (\mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathrm{Prov}(\ulcorner D \urcorner) \to \mathrm{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathrm{Prov}(\ulcorner A \urcorner) \to A$	Assumption
8.	$\mathrm{Prov}(\ulcorner D \urcorner) \to A$	PC: 6, 7
9.	D	PC: 1, 8
10.	$\mathrm{Prov}(\ulcorner D \urcorner)$	D1 from 9
11.	A	PC: 8, 10

1.	$D \leftrightarrow (\mathsf{Prov}(\ulcorner D \urcorner) \to A)$	FPT
$\vdots$	$\vdots$	$\vdots$
4.	$\mathsf{Prov}(\ulcorner D \urcorner) \to (\mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner))$	PC: 2, 3
5.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner \mathsf{Prov}(\ulcorner D \urcorner) \urcorner)$	D3
6.	$\mathsf{Prov}(\ulcorner D \urcorner) \to \mathsf{Prov}(\ulcorner A \urcorner)$	PC: 4, 5
7.	$\mathsf{Prov}(\ulcorner A \urcorner) \to A$	Assumption
8.	$\mathsf{Prov}(\ulcorner D \urcorner) \to A$	PC: 6, 7
9.	D	PC: 1, 8
10.	$\mathsf{Prov}(\ulcorner D \urcorner)$	D1 from 9
11.	A	PC: 8, 10

‘**PA** couldn’t be more modest about its own veracity’

By Löb’s Theorem, it is not true that for all sentences φ,

$$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$$

‘**PA** couldn’t be more modest about its own veracity’

By Löb’s Theorem, it is not true that for all sentences φ,

$$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$$

Statement	It is not true that...
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner)$ implies $\mathbf{PA} \vdash \varphi$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi$

‘**PA** couldn’t be more modest about its own veracity’

By Löb’s Theorem, it is not true that for all sentences φ,

$$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$$

Statement	It is not true that...
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner)$ implies $\mathbf{PA} \vdash \varphi$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi$
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\varphi \urcorner)$ implies $\mathbf{PA} \nvdash \mathrm{Prov}(\ulcorner \varphi \urcorner)$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\varphi \urcorner) \to \neg\mathrm{Prov}(\ulcorner \varphi \urcorner)$

‘**PA** couldn't be more modest about its own veracity’

By Löb's Theorem, it is not true that for all sentences φ,

$$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$$

Statement	It is not true that...
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner)$ implies $\mathbf{PA} \vdash \varphi$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \varphi \urcorner) \to \varphi$
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\varphi \urcorner)$ implies $\mathbf{PA} \nvdash \mathrm{Prov}(\ulcorner \varphi \urcorner)$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\varphi \urcorner) \to \neg\mathrm{Prov}(\ulcorner \varphi \urcorner)$
$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\mathrm{Prov}(\ulcorner \varphi \urcorner) \urcorner)$ implies $\mathbf{PA} \vdash \neg\mathrm{Prov}(\varphi)$	$\mathbf{PA} \vdash \mathrm{Prov}(\ulcorner \neg\mathrm{Prov}(\ulcorner \varphi \urcorner) \urcorner) \to \neg\mathrm{Prov}(\ulcorner \varphi \urcorner)$

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
- ▶ A Primer on Epistemic and Doxastic Logic
- ▶ Anti-Expert Paradoxes
- ▶ The Knower Paradox and variants
- ▶ Epistemic Arithmetic
- ▶ Gödel's Disjunction

Rineke Verbrugge (2024). *Provability Logic*. The Stanford Encyclopedia of Philosophy (Summer 2024 Edition), Edward N. Zalta & Uri Nodelman (eds.), https://plato.stanford.edu/archives/sum2024/entries/logic-provability/.

Propositional Modal Logic

$$p \mid \neg\varphi \wedge \varphi \wedge \psi \mid \Box\varphi$$

where $p \in \mathsf{AT}$ (at set of atomic propositions).

The intended interpretation of $\Box\varphi$ is "there is a proof (in **PA**) of φ".

Propositional Modal Logic

$$p \mid \neg\varphi \wedge \varphi \wedge \psi \mid \Box\varphi$$

where $p \in \mathrm{AT}$ (at set of atomic propositions).

The intended interpretation of $\Box\varphi$ is "there is a proof (in **PA**) of φ".

A **frame** is a tuple (W, R) such that $W \neq \emptyset$ and $R \subseteq W \times W$.

Propositional Modal Logic

$$p \mid \neg\varphi \wedge \varphi \wedge \psi \mid \Box\varphi$$

where $p \in \mathsf{AT}$ (at set of atomic propositions).

The intended interpretation of $\Box\varphi$ is "there is a proof (in **PA**) of φ".

A **frame** is a tuple (W, R) such that $W \neq \emptyset$ and $R \subseteq W \times W$.

A **model** is a tuple (W, R, V) where (W, R) is a frame and $V : \mathsf{AT} \to \wp(W)$.

Truth/Validity

For a model $\mathcal{M} = (W, R, V)$ and $w \in W$, we write $\mathcal{M} \models \varphi$ when φ is true at w in $\mathcal{M}$.

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff $\mathcal{M}, w \not\models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$
- $\mathcal{M}, w \models \Box\varphi$ iff for all $v \in W$, if $w \, R \, v$, then $\mathcal{M}, v \models \varphi$

For a frame $\mathcal{F} = (W, R)$, φ is **valid on** $\mathcal{F}$, denoted $\mathcal{F} \models \varphi$, when $\mathcal{M}, w \models \varphi$ for all models $\mathcal{M}$ based on $\mathcal{F}$ and $w \in W$.

Provability Logic: **GL**

K	$\Box(\varphi \to \psi) \to (\Box\varphi \to \psi)$
L	$\Box(\Box\varphi \to \varphi) \to \Box\varphi$
MP	$\varphi, \varphi \to \psi \ \therefore\ \psi$
NEC	$\varphi \ \therefore\ \Box\varphi$

Some Results

- $\mathbf{GL} \vdash \Box\varphi \to \Box\Box\varphi$.

Some Results

- $\mathbf{GL} \vdash \Box\varphi \to \Box\Box\varphi$.
- $\Box(\Box\varphi \to \varphi) \to \Box\varphi$ is valid on a frame (W, R) if, and only if, R is transitive and converse well-founded (there are no infinite ascending sequences, that is sequences of the form $w_1 \; R \; w_2 \; R \; w_3 \cdots$).

Some Results

- $\mathbf{GL} \vdash \Box\varphi \to \Box\Box\varphi$.
- $\Box(\Box\varphi \to \varphi) \to \Box\varphi$ is valid on a frame (W, R) if, and only if, R is transitive and converse well-founded (there are no infinite ascending sequences, that is sequences of the form $w_1\ R\ w_2\ R\ w_3 \cdots$).
- The logic **GL** is not compact:

$$\Gamma = \{\Diamond p_0, \Box(p_0 \to \Diamond p_1), \Box(p_1 \to \Diamond p_2), \ldots, \Box(p_n \to \Diamond p_{n+1}), \ldots\}.$$

is finitely satisfiable, but not satisfiable.

Some Results

- $\mathbf{GL} \vdash \Box\varphi \to \Box\Box\varphi$.
- $\Box(\Box\varphi \to \varphi) \to \Box\varphi$ is valid on a frame (W, R) if, and only if, R is transitive and converse well-founded (there are no infinite ascending sequences, that is sequences of the form $w_1 \; R \; w_2 \; R \; w_3 \cdots$).
- The logic **GL** is not compact:

$$\Gamma = \{\Diamond p_0, \Box(p_0 \to \Diamond p_1), \Box(p_1 \to \Diamond p_2), \ldots, \Box(p_n \to \Diamond p_{n+1}), \ldots\}.$$

  is finitely satisfiable, but not satisfiable.
- The logic **GL** is sound and weakly complete with respect to the class of frames that are transitive and converse well-founded.

Arithmetic Completeness

An **arithmetic translation** is a function t such that

1. For all $p \in \mathrm{At}$, $t(p)$ is a sentence of $\mathcal{L}_A$
2. t commutes with the boolean connectives: $t(\neg\varphi) = \neg t(\varphi)$, $t(\varphi \wedge \psi) = t(\varphi) \wedge t(\psi)$, etc.
3. $t(\Box\varphi) = \mathrm{Prov}_{\mathbf{PA}}(\ulcorner t(\varphi) \urcorner)$

Theorem (Solovay 1976).

$$\mathbf{GL} \vdash \varphi \text{ iff for every arithmetic translation } t,\ \mathbf{PA} \vdash t(\varphi).$$

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
- ▶ A Primer on Epistemic and Doxastic Logic
- ▶ Anti-Expert Paradoxes
- ▶ The Knower Paradox and variants
- ▶ Epistemic Arithmetic
- ▶ Gödel's Disjunction

Predicate vs. Operator Approach to Modality

Predicate Approach	'$2 + 2 = 4$' is necessary
Operator Approach	It is necessary that $2 + 2 = 4$.

Predicate vs. Operator Approach to Modality

Predicate Approach	'$2+2=4$' is necessary
Operator Approach	It is necessary that $2+2=4$.

- We have treated 'provability' both as a predicate ($\mathrm{Prov}_{\mathbf{T}}(\cdot)$) and as a sentential operator (in **GL**)
- Truth is typically only treated as a predicate

Predicate vs. Operator Approach to Modality

Whether necessity, knowledge, belief, future and past truth, obligation, and other modalitities should be formalised by operators or by predicates was a matter of dispute up to the early sixties between two almost equally strong parties. Then two technical achievements helped the operator approach to an almost complete triumph over the predicate approach that had been advocated by illustrious philosophers like Quine. (p. 180)

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.

Predicate vs. Operator Approach to Modality

1. 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.

Predicate vs. Operator Approach to Modality

1. 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.

2. 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.

Theorem (Tarski/Gödel). Let $\mathbf{T}$ be a theory extending $\mathbf{Q}$ and T a unary predicate such that for all sentences φ:

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

Then, $\mathbf{T}$ is inconsistent.

Proof. By the Fixed Point Theorem, there is a sentence D such that

$$\mathbf{T} \vdash D \leftrightarrow \neg T(\ulcorner D \urcorner)$$

Theorem (Tarski/Gödel). Let $\mathbf{T}$ be a theory extending $\mathbf{Q}$ and T a unary predicate such that for all sentences φ:

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

Then, $\mathbf{T}$ is inconsistent.

Proof. By the Fixed Point Theorem, there is a sentence D such that

$$\mathbf{T} \vdash D \leftrightarrow \neg T(\ulcorner D \urcorner)$$

But, since $\mathbf{T} \vdash T(\ulcorner D \urcorner) \leftrightarrow D$, the contradiction is immediate.

Montague's Theorem

Theorem (Montague (1963))

Suppose **T** is a theory and $\Box(x)$ is a formula such that for all sentences φ,

(T) $\quad \mathbf{T} \vdash \Box(\ulcorner\varphi\urcorner) \to \varphi$

(Nec) If $\mathbf{T} \vdash \varphi$, then $\mathbf{T} \vdash \Box(\ulcorner\varphi\urcorner)$

(Q) $\quad \mathbf{Q} \subseteq \mathbf{T}$

Then **T** is inconsistent.

R. Montague (1963). *Syntactical Treatment of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability*. Acta Philosophica Finnica, 16, pp. 153 - 167.

1. $D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$ FPT (using Q)

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth
3.	$\Box(\ulcorner D \urcorner) \rightarrow \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth
3.	$\Box(\ulcorner D \urcorner) \rightarrow \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2
4.	$\neg\Box(\ulcorner D \urcorner)$	PC: 3

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \to D$	Truth
3.	$\Box(\ulcorner D \urcorner) \to \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2
4.	$\neg\Box(\ulcorner D \urcorner)$	PC: 3
5.	D	PC: 1, 4

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth
3.	$\Box(\ulcorner D \urcorner) \rightarrow \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2
4.	$\neg\Box(\ulcorner D \urcorner)$	PC: 3
5.	D	PC: 1, 4
6.	$\Box(\ulcorner D \urcorner)$	Nec: 5

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth
3.	$\Box(\ulcorner D \urcorner) \rightarrow \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2
4.	$\neg\Box(\ulcorner D \urcorner)$	PC: 3
5.	D	PC: 1, 4
6.	$\Box(\ulcorner D \urcorner)$	Nec: 5
7.	$\perp$	3, 6

1.	$D \leftrightarrow \neg\Box(\ulcorner D \urcorner)$	FPT (using Q)
2.	$\Box(\ulcorner D \urcorner) \rightarrow D$	Truth
3.	$\Box(\ulcorner D \urcorner) \rightarrow \neg\Box(\ulcorner D \urcorner)$	PC: 1, 2
4.	$\neg\Box(\ulcorner D \urcorner)$	PC: 3
5.	D	PC: 1, 4
6.	$\Box(\ulcorner D \urcorner)$	Nec: 5
7.	$\bot$	3, 6

A Problem with the Operator Approach

The operator approach suffers from a severe drawback: it restricts the expressive power of the language in a dramatic way because it rules out quantification in the following sense:

There is no direct formalisation of a sentence like

"All tautologies of propositional logic are necessary."

- Substitutional quantification: $\forall A(P(A) \to \Box A)$, where P is a predicate and $\Box$ is an operator.

- Substitutional quantification: $\forall A(P(A) \to \Box A)$, where P is a predicate and $\Box$ is an operator. However, this quantification does not come with a semantics, only rules and axioms. Also, why are the following sentences formalized using different types of quantification?
  - "All Σ_1 sentences are provable"
  - "All Σ_1 sentences are necessary"

- Substitutional quantification: $\forall A(P(A) \rightarrow \Box A)$, where P is a predicate and $\Box$ is an operator. However, this quantification does not come with a semantics, only rules and axioms. Also, why are the following sentences formalized using different types of quantification?
  - "All Σ_1 sentences are provable"
  - "All Σ_1 sentences are necessary"

- Rather than "x is necessary", say "x is necessarily true". Thus, $\Box x$ is replaced by $\Box Tx$, where T is a truth predicate.

- Substitutional quantification: $\forall A(P(A) \to \Box A)$, where P is a predicate and $\Box$ is an operator. However, this quantification does not come with a semantics, only rules and axioms. Also, why are the following sentences formalized using different types of quantification?
  - "All Σ_1 sentences are provable"
  - "All Σ_1 sentences are necessary"

- Rather than "x is necessary", say "x is necessarily true". Thus, $\Box x$ is replaced by $\Box Tx$, where T is a truth predicate. However, why should truth and necessity be treated differently at the syntactic level; and, this would mean that the theory of necessity would inherent all the semantical paradoxes.

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.

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

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

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 $V(w) \models \Box\ulcorner A \urcorner$.
- $V(w) \models \Box(\ulcorner A \to B \urcorner) \to (\Box\ulcorner A \urcorner \to \Box\ulcorner B \urcorner)$

$$\forall x \forall y ((\mathrm{Sent}(x) \land \mathrm{Sent}(y)) \rightarrow (\Box \ulcorner x \mathbin{\dot{\rightarrow}} y \urcorner \rightarrow (\Box x \rightarrow \Box y)))$$

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.

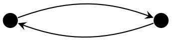
- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A \urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A \urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.
- Thus, $V(w) \models A$.
- Hence, $V(w) \models \neg\Box\ulcorner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

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.

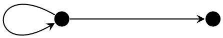
- We have $\mathbf{PA} \vdash A \leftrightarrow \neg\Box\ulcorner A\urcorner$, and so it holds at every world.
- If $V(w) \models \neg A$, then $V(w) \models \Box\ulcorner A\urcorner$.
- So, by reflexivity, $V(w) \models A$. Contradiction.
- Thus, $V(w) \models A$.
- Hence, $V(w) \models \neg\Box\ulcorner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



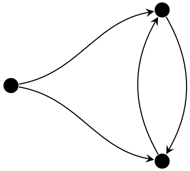
   Use the fixed point: $A \leftrightarrow \neg\Box\ulcorner\Box\ulcorner A\urcorner\urcorner$

2. The following frame does not admit a valuation:



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

3. The following frame does not admit a valuation:



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

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

$$0 \longrightarrow 1 \longrightarrow 2 \longrightarrow \cdots$$

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,

$$\Box\ulcorner A\urcorner \to \Box\ulcorner\Box\ulcorner A\urcorner\urcorner$$

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:

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

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

Fact. In 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:

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

Predicate Approaches to Modality

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