Epistemic Arithmetic

Eric Pacuit
University of Maryland

Lecture 4, ESSLLI 2025

July 31, 2025

The Knower Paradox

Theorem (Montague-Kaplan 1960)

Let **T** be an axiomatizable extension of **Q**, with $I(x, y)$ a formula of expressing derivability between sentences in **T**, and K a (perhaps complex) unary predicate satisfying, for all sentences φ and ψ:

(T) $K(\varphi) \to \varphi$

(U) $K(K\varphi \to \varphi)$

(I) $(K(\varphi) \wedge I(\varphi, \psi)) \to K(\psi)$

then **T** is inconsistent.

1.	$D \leftrightarrow K(\neg D)$	FPT (using Q)
2.	$K(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5.	$I(K(\neg D) \rightarrow \neg D, \neg D)$	2-4
6.	$K(K(\neg D) \rightarrow \neg D)$	U
7.	$(K(K(\neg D) \rightarrow \neg D) \wedge I(K(\neg D) \rightarrow \neg D, \neg D)) \rightarrow K(\neg D)$	I
8.	$K(\neg D)$	PC, MP: 5 & 6, 7
9.	D	PC: 1, 8
10.	$\perp$	4, 9

How should we solve this paradox? Should knowledge entail truth? Should we accept the epistemic closure principle or not? Should the syntax be changed in such a way that statements that lead to paradoxes are eliminated?

Theorem (Koons, Turner)

Let **T** be a theory extending **Q**, with B a (perhaps complex) unary predicate, such that **T** satisfies, for all sentences φ and ψ:

(4) $B(\varphi) \rightarrow B(B(\varphi))$

(D) $B(\neg\varphi) \rightarrow \neg B(\varphi)$

(Nec) If $\mathbf{T} \vdash \varphi$, then $\mathbf{T} \vdash B(\varphi)$

(Re) If $\mathbf{T} \vdash \varphi \leftrightarrow \psi$, then $\mathbf{T} \vdash B(\varphi) \leftrightarrow B(\psi)$

then T is inconsistent.

Theorem (Cross 2001)

Let $\mathbf{T}$ be an axiomatizable theory extending $\mathbf{Q}$, with K a (perhaps complex) predicate. Let $K'(x)$ be the predicate defined by the formula:

$$\exists y(K(y) \wedge I(y, x))$$

where $I(y, x)$ is a predicate expressing derivability between sentences in $\mathbf{T}$. Suppose $\mathbf{T}$ satisfies the following axiom schemata:

(T′) $K'(\varphi) \rightarrow \varphi$

(U′) $K'(K'(\varphi) \rightarrow \varphi)$

then $\mathbf{T}$ is inconsistent.

Theorem (Cross's 'Knowledge-Plus Knower')

Let $\mathbf{T}$ be an axiomatizable theory extending $\mathbf{Q}$, with K and K' defined as previously, and such that $\mathbf{T}$ satisfies, for every sentence φ:

(T') $\quad K'(\varphi) \rightarrow \varphi$

(U^{+}) $\quad K(K'(\varphi) \rightarrow \varphi)$

then $\mathbf{T}$ is inconsistent.

Anderson's Solution

C. Anthony Anderson (1983). *The Paradox of the Knower*. The Journal of Philosophy, 80, 6, pp. 338-355.

Anderson's Solution

$\mathcal{L}_0$: the smallest extension of $\mathcal{L}_A$ such that if $\varphi, \psi \in \mathcal{L}_A$, then $K_0(\varphi), I_0(\varphi, \psi) \in \mathcal{L}_0$, closed under Boolean operators.

Anderson's Solution

$\mathcal{L}_0$: the smallest extension of $\mathcal{L}_A$ such that if $\varphi, \psi \in \mathcal{L}_A$, then $K_0(\varphi), I_0(\varphi, \psi) \in \mathcal{L}_0$, closed under Boolean operators.

$\mathcal{L}_{i+1}$: the smallest extension of $\mathcal{L}_i$ such that if $\varphi, \psi \in \mathcal{L}_i$, then $K_{i+1}(\varphi), I_{i+1}(\varphi, \psi) \in \mathcal{L}_{i+1}$, closed under Boolean operators.

Anderson's Solution

$\mathcal{L}_0$: the smallest extension of $\mathcal{L}_A$ such that if $\varphi, \psi \in \mathcal{L}_A$, then $K_0(\varphi), I_0(\varphi, \psi) \in \mathcal{L}_0$, closed under Boolean operators.

$\mathcal{L}_{i+1}$: the smallest extension of $\mathcal{L}_i$ such that if $\varphi, \psi \in \mathcal{L}_i$, then $K_{i+1}(\varphi), I_{i+1}(\varphi, \psi) \in \mathcal{L}_{i+1}$, closed under Boolean operators.

$\mathcal{L}_\omega$: $\bigcup_{i \in \omega} \mathcal{L}_i$

Anderson's Solution

$\mathcal{L}_0$: the smallest extension of $\mathcal{L}_A$ such that if $\varphi, \psi \in \mathcal{L}_A$, then $K_0(\varphi), I_0(\varphi, \psi) \in \mathcal{L}_0$, closed under Boolean operators.

$\mathcal{L}_{i+1}$: the smallest extension of $\mathcal{L}_i$ such that if $\varphi, \psi \in \mathcal{L}_i$, then $K_{i+1}(\varphi), I_{i+1}(\varphi, \psi) \in \mathcal{L}_{i+1}$, closed under Boolean operators.

$\mathcal{L}_\omega$: $\bigcup_{i \in \omega} \mathcal{L}_i$

K_i indicates a certain level of knowledge. Anderson gives an "intuitive motivation": Some sentence that cannot be in a set of statements known at level i can still be provable. By understanding the proof of such a statement, one knows this sentence at level $i + 1$.

Anderson's Solution

$gn(\mathcal{L}_\omega) = \{gn(\alpha) \mid \alpha \in \mathcal{L}_\omega\}$ is the set of Gödel numbers of each formula in $\mathcal{L}_\omega$. Suppose that V_p is an interpretation of $\mathcal{L}_A$:

- V_0 extends V_p to $\mathcal{L}_0$
- V_{i+1} extends V_i to $\mathcal{L}_{i+1}$
- $V_i(K_i) \subseteq gn(\mathcal{L}_\omega)$
- $V_i(I_i) \subseteq gn(\mathcal{L}_\omega) \times gn(\mathcal{L}_\omega)$
- $V = \bigcup_{i \in \omega} V_i$

Anderson's Solution

$$
\begin{aligned}
\mathbf{T}_0 &= \mathbf{Q} \cup \{K_0(\ulcorner\varphi\urcorner) \to \varphi \mid \varphi \in \mathcal{L}_\omega\} \\
\mathbf{T}_{i+1} &= \mathbf{T}_i \cup \{K_{i+1}(\ulcorner\varphi\urcorner) \to \varphi \mid \varphi \in \mathcal{L}_\omega\} \\
V_0(K_0(\ulcorner\varphi\urcorner)) &= 1 \text{ if and only if } \mathbf{Q} \vdash \varphi \\
V_{i+1}(K_{i+1}(\ulcorner\varphi\urcorner)) &= 1 \text{ if and only if } \mathbf{T}_i \vdash \varphi \\
V_0(I_0(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) &= 1 \text{ if and only if } \mathbf{Q} \vdash \varphi \to \psi \\
V_{i+1}(I_{i+1}(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) &= 1 \text{ if and only if } \mathbf{T}_i \vdash \varphi \to \psi \\
\mathbf{T}_\omega &= \textstyle\bigcup_{i\in\omega} \mathbf{T}_i.
\end{aligned}
$$

Anderson's Solution

- $V_i(K_i) \subseteq V_{i+1}(K_{i+1})$.
- $V_i(I_i) \subseteq V_{i+1}(I_{i+1})$.
- If $n = gn(\varphi) \in V_i(K_i)$, then $\exists j \geq i$ such that $V_j(\varphi) = 1$.
- If $n = gn(\varphi), m = gn(\psi), (n, m) \in V_i(I_i)$, then $\exists j \geq i$ such that $V_j(\varphi \to \psi) = 1$.
- If $(n, m) \in V_i(I_i), n \in V_i(K_i)$, then $m \in V_i(K_i)$.

Anderson's Solution

$$
\begin{aligned}
V(K_i(\ulcorner\varphi\urcorner) \to \varphi) &= 1 \\
V([I_i(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge K_i(\ulcorner\varphi\urcorner)] \to K_i(\ulcorner\psi\urcorner)) &= 1 \\
V(K_{i+1}(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner)) &= 1
\end{aligned}
$$

Anderson's Solution

$$
\begin{aligned}
V(K_i(\ulcorner\varphi\urcorner) \to \varphi) &= 1 \\
V([I_i(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge K_i(\ulcorner\varphi\urcorner)] \to K_i(\ulcorner\psi\urcorner)) &= 1 \\
V(K_{i+1}(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner)) &= 1
\end{aligned}
$$

$$
K_{i+1}(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner) \quad \textit{vs.} \quad K_i(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner)
$$

Anderson's Solution

$$
\begin{aligned}
V(K_i(\ulcorner\varphi\urcorner) \to \varphi) &= 1 \\
V([I_i(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \wedge K_i(\ulcorner\varphi\urcorner)] \to K_i(\ulcorner\psi\urcorner)) &= 1 \\
V(K_{i+1}(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner)) &= 1
\end{aligned}
$$

$$
K_{i+1}(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner) \quad \textit{vs.} \quad K_i(\ulcorner K_i(\ulcorner\varphi\urcorner) \to \varphi\urcorner)
$$

$$
K_i(\ulcorner\varphi\urcorner) \to K_j(\ulcorner\varphi\urcorner) \text{ for } j \geq i.
$$

$$
I_i(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \to I_j(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner) \text{ for } j \geq i.
$$

Blocking the Knower Paradox

1.	$D \leftrightarrow K(\neg D)$	FPT
2.	$K(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5.	$I(K(\neg D) \rightarrow \neg D, \neg D)$	2-4
6.	$K(K(\neg D) \rightarrow \neg D)$	U
7.	$(K(K(\neg D) \rightarrow \neg D) \wedge I(K(\neg D) \rightarrow \neg D, \neg D)) \rightarrow K(\neg D)$	I
8.	$K(\neg D)$	PC: 5, 6, 7
9.	D	PC: 1, 8
10.	$\bot$	4, 9

Blocking the Knower Paradox

1.	$D \leftrightarrow K_i(\neg D)$	FPT
2.	$K_i(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5.	$I_i(K_i(\neg D) \rightarrow \neg D, \neg D)$	2-4

Blocking the Knower Paradox

1.	$D \leftrightarrow K_i(\neg D)$	FPT
2.	$K_i(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5′.	$I_{i+1}(K_i(\neg D) \rightarrow \neg D, \neg D)$	2-4

Blocking the Knower Paradox

1.	$D \leftrightarrow K_i(\neg D)$	FPT
2.	$K_i(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5′.	$I_{i+1}(K_i(\neg D) \rightarrow \neg D, \neg D)$	2-4
6.	$K_{i+1}(K_i(\neg D) \rightarrow \neg D)$	
7.	$(K_{i+1}(K_i(\neg D) \rightarrow \neg D) \wedge I_{i+1}(K_i(\neg D) \rightarrow \neg D, \neg D)) \rightarrow K_{i+1}(\neg D)$	I
8.	$K_{i+1}(\neg D)$	PC: 5, 6, 7

Blocking the Knower Paradox

1.	$D \leftrightarrow K_i(\neg D)$	FPT
2.	$K_i(\neg D) \rightarrow \neg D$	Truth
3.	$D \rightarrow \neg D$	PC: 1, 2
4.	$\neg D$	PC: 3
5′.	$I_{i+1}(K_i(\neg D) \rightarrow \neg D, \neg D)$	2-4
6.	$K_{i+1}(K_i(\neg D) \rightarrow \neg D)$	
7.	$(K_{i+1}(K_i(\neg D) \rightarrow \neg D) \wedge I_{i+1}(K_i(\neg D) \rightarrow \neg D, \neg D)) \rightarrow K_{i+1}(\neg D)$	I
8.	$K_{i+1}(\neg D)$	PC: 5′, 6, 7
9.	D	PC: 1, 8
10.	$\bot$	4, 9

Solutions to the Knower Paradox

Paul Égré (2005). *The Knower Paradox in the Light of Provability Interpretations of Modal*. Journal of Logic, Language and Information, 14, pp. 13 - 48.

Solutions to the Knower Paradox

Paul Égré (2005). *The Knower Paradox in the Light of Provability Interpretations of Modal*. Journal of Logic, Language and Information, 14, pp. 13 - 48.

Francesa Poggiolesi (2007). *Three Different Solutions to the Knower Paradox*. Annali del Dipartimento di Filosofia, 13(1), pp. 147 - 163.

Solutions to the Knower Paradox

Paul Égré (2005). *The Knower Paradox in the Light of Provability Interpretations of Modal*. Journal of Logic, Language and Information, 14, pp. 13 - 48.

Francesa Poggiolesi (2007). *Three Different Solutions to the Knower Paradox*. Annali del Dipartimento di Filosofia, 13(1), pp. 147 - 163.

Mirjam de Vos, Rineke Verbrugge, and Barteld Kooi (2023). *Solutions to the Knower Paradox in the Light of Haack's Criteria*. Journal of Philosophical Logic, 52, pp. 1101 - 1132.

Knower Paradox in the Quantified Logic of Proofs

W. Dean (2014). *Montague's paradox, informal provability, and explicit modal logic*. Notre Dame Journal of Formal Logic, 55(2), pp. 157 - 196.

W. Dean and H. Kurokawa (2014). *The paradox of the Knower revisited*. Annals of Pure and Applied Logic, 165(1), pp. 199 - 224.

$K(\ulcorner A \urcorner)$ if and only if there exists a proof p which demonstrates A

"...[W]e do not, at least *ipso facto*, wish to suggest that $K(x)$ must be analyzed in terms of provability in a specific axiom system in the manner in which Cross's definition of $K'(x)$ might seem to suggest. For we might also adopt the view that in p should be understood as ranging over some class of informal proofs—i.e. intuitively correct pieces of reasoning which carry conviction for the epistemic subject in question." (Dean and Kurokawa, p. 11)

Logic of Proofs

S. Artemov and M. Fitting (2024). *Justification Logic*. Stanford Encyclopedia of Philosophy, http://plato.stanford.edu/entries/logic-justification/.

S. Artemov and M. Fitting. *Justification Logic: Reasoning with Reasons*. Cambridge University Press, 2019.

R. Kuznets and T. Studer (2019). *Logics of Proofs and Justifications*. College Publications.

$$t := c \mid x \mid t + s \mid !t \mid t \cdot s$$

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \to \psi \mid t : \varphi$$

$$t := c \mid x \mid t + s \mid !t \mid t \cdot s$$

$$\varphi := p \mid \neg\varphi \mid \varphi \wedge \psi \mid \varphi \vee \psi \mid \varphi \to \psi \mid t : \varphi$$

- $t : \varphi$ means "t is a justification/proof of φ", or "φ is so for reason t".
- x is a variable: an arbitrary justification
- c is a constant: justifications of formulas we do not analyze further (axioms)
- $t \cdot u$ justifies φ whenevery u justifies some formula ψ and t justifies $\psi \to \varphi$
- if t justifies φ, $!t$ justifies that fact.

LP

Factivity: $t : \phi \to \phi$

Application: $t : (\phi \to \psi) \to (s : \phi \to t \cdot s : \psi)$

Proof checker $t : \phi \to !t : t : \phi$

Constant specification: a set of formulas of the form $c_1 : c_2 : \cdots c_n : \varphi$ where φ is an instance of the axom from the list above, $n \geq 0$ and $c_1, \ldots, c_n$ are justification constants.

$$\vdash_{S4} (\Box P \vee \Box Q) \to \Box(\Box P \vee \Box Q)$$

$$\vdash_{LP} (x : P \vee y : Q) \to [a{\cdot}!x + b{\cdot}!y] : (x : P \vee y : Q)$$

Quantified Logic of Proofs

- $(\forall x)\,(\forall x)\,((x:\varphi \to \psi \wedge y:\varphi) \to x \cdot y:\psi)$
- $(\exists x)\,(x:\varphi \to \varphi)$
- $(\exists y)\,y:((\exists x)\,(x:\varphi \to \varphi))$

M. Fitting (2008). *A quantified logic of evidence*. Annals of Pure and Applied Logic, 152(1-3), pp. 67-83.

The Knower Paradox in the QLP

1. $d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$ Factivity, MP

The Knower Paradox in the QLP

1. $d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$ Factivity, MP
2. $d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$ Derivation in QLP

The Knower Paradox in the QLP

1.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$	Factivity, MP
2.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$	Derivation in QLP
3.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \neg D$	1, 2

The Knower Paradox in the QLP

1.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$	Factivity, MP
2.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$	Derivation in QLP
3.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \neg D$	1, 2
4.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash t(d) : \neg D$	for some $t(d)$ by Internalization

The Knower Paradox in the QLP

1.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$	Factivity, MP
2.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$	Derivation in QLP
3.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \neg D$	1, 2
4.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash t(d) : \neg D$	for some $t(d)$ by Internalization
5.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D$	from 4, by $\exists$-intro

The Knower Paradox in the QLP

1.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$	Factivity, MP
2.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$	Derivation in QLP
3.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \neg D$	1, 2
4.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash t(d) : \neg D$	for some $t(d)$ by Internalization
5.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D$	from 4, by $\exists$-intro
6.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D$	from 1, 5

The Knower Paradox in the QLP

1.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D \leftrightarrow (\exists x)\, x : \neg D$	Factivity, MP
2.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D \rightarrow \neg D$	Derivation in QLP
3.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \neg D$	1, 2
4.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash t(d) : \neg D$	for some $t(d)$ by Internalization
5.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash (\exists x)\, x : \neg D$	from 4, by $\exists$-intro
6.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash D$	from 1, 5
7.	$d : (D \leftrightarrow (\exists x)\, x : \neg D) \vdash \bot$	from 3, 6

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

Doxastic Logic: Models

Model: $\langle W, R, V \rangle$

States/possible worlds: $W \neq \emptyset$

Quasi-partitions: $R \subseteq W \times W$ is serial, transitive and Euclidean

Doxastic Logic: Models

Model: $\langle W, R, V \rangle$

States/possible worlds: $W \neq \emptyset$

Quasi-partitions: $R \subseteq W \times W$ is serial, transitive and Euclidean

- *serial*: for all $w \in W$, there is a $v \in W$ such that $w\,R\,v$
- *transitive*: for all $w, v, u \in W$, if $w\,R\,v$ and $v\,R\,u$, then $w\,R\,u$
- *Euclidean*: for all $w, v, u \in W$, if $w\,R\,v$ and $w\,R\,u$, then $v\,R\,u$

Doxastic Logic: Models

Model: $\langle W, R, V \rangle$

States/possible worlds: $W \neq \emptyset$

Quasi-partitions: $R \subseteq W \times W$ is serial, transitive and Euclidean

- *serial*: for all $w \in W$, there is a $v \in W$ such that $w\,R\,v$
- *transitive*: for all $w, v, u \in W$, if $w\,R\,v$ and $v\,R\,u$, then $w\,R\,u$
- *Euclidean*: for all $w, v, u \in W$, if $w\,R\,v$ and $w\,R\,u$, then $v\,R\,u$

Valuation function: $V : \text{At} \to \wp(W)$, where At is a set of atomic propositions.

Doxastic Logic: Language and Semantics

$$p \mid \varphi \wedge \varphi \mid \neg\varphi \mid B\varphi$$

Doxastic Logic: Language and Semantics

$$p \mid \varphi \wedge \varphi \mid \neg\varphi \mid B\varphi$$

Boolean connectives:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff it is not the case that $\mathcal{M}, w \models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

Doxastic Logic: Language and Semantics

$$p \mid \varphi \wedge \varphi \mid \neg\varphi \mid B\varphi$$

Boolean connectives:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff it is not the case that $\mathcal{M}, w \models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

Belief operators: $\mathcal{M}, w \models B\varphi$ iff for all v, if $w\,R\,v$, then $\mathcal{M}, v \models \varphi$.

Doxastic Logic: Language and Semantics

$$p \mid \varphi \wedge \varphi \mid \neg\varphi \mid B\varphi$$

Boolean connectives:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff it is not the case that $\mathcal{M}, w \models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

Belief operators: $\mathcal{M}, w \models B\varphi$ iff for all v, if $w\,R\,v$, then $\mathcal{M}, v \models \varphi$.

$$\mathcal{M}, w \models B\varphi \text{ iff } R(w) \subseteq [\![\varphi]\!]^{\mathcal{M}}$$

Doxastic Logic: Language and Semantics

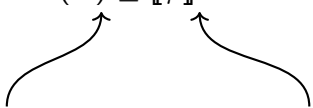
$$p \mid \varphi \wedge \varphi \mid \neg\varphi \mid B\varphi$$

Boolean connectives:

- $\mathcal{M}, w \models p$ iff $w \in V(p)$
- $\mathcal{M}, w \models \neg\varphi$ iff it is not the case that $\mathcal{M}, w \models \varphi$
- $\mathcal{M}, w \models \varphi \wedge \psi$ iff $\mathcal{M}, w \models \varphi$ and $\mathcal{M}, w \models \psi$

Belief operators: $\mathcal{M}, w \models B\varphi$ iff for all v, if $w\,R\,v$, then $\mathcal{M}, v \models \varphi$.

$$\mathcal{M}, w \models B\varphi \text{ iff } R(w) \subseteq [\![\varphi]\!]^{\mathcal{M}}$$



$$\{v \mid w\,R\,v\} \qquad \{v \mid \mathcal{M}, w \models \varphi\}$$

Doxastic Logic: **KD45**

K	$B(\varphi \to \psi) \to (B\varphi \to B\psi)$
D	$B\varphi \to \neg B\neg\varphi$
4	$B\varphi \to BB\varphi$
5	$\neg B\varphi \to B\neg B\varphi$

Doxastic Logic: **KD45**

K	$B(\varphi \to \psi) \to (B\varphi \to B\psi)$
D	$B\varphi \to \neg B\neg\varphi$
4	$B\varphi \to BB\varphi$
5	$\neg B\varphi \to B\neg B\varphi$

The logic **KD45** adds the above axiom schemes to an axiomatization of classical propositional logic with the rules Modus Ponens, Substitution of Equivalents, and Necessitation (from φ infer $B\varphi$).

KD45 is sound and strongly complete with respect to all quasi-partition frames.

Exercise: Show that the following axiom schemes and rules are valid on quasi-partition models and are theorems of **KD45**:

- agglomeration: $(B\varphi \wedge B\psi) \rightarrow B(\varphi \wedge \psi)$
- consistency: $\neg B\bot$
- monotonicity: From $\varphi \rightarrow \psi$ infer $B\varphi \rightarrow B\psi$

Exercise: Show that the following axiom schemes and rules are valid on quasi-partition models and are theorems of **KD45**:

- agglomeration: $(B\varphi \wedge B\psi) \rightarrow B(\varphi \wedge \psi)$
- consistency: $\neg B\bot$
- monotonicity: From $\varphi \rightarrow \psi$ infer $B\varphi \rightarrow B\psi$
- secondary-reflexivity: for all $w, v \in W$, if $w\,R\,v$ then $v\,R\,v$
  $B(B\varphi \rightarrow \varphi)$

Exercise: Show that the following axiom schemes and rules are valid on quasi-partition models and are theorems of **KD45**:

- agglomeration: $(B\varphi \wedge B\psi) \rightarrow B(\varphi \wedge \psi)$
- consistency: $\neg B\bot$
- monotonicity: From $\varphi \rightarrow \psi$ infer $B\varphi \rightarrow B\psi$
- secondary-reflexivity: for all $w, v \in W$, if $w\,R\,v$ then $v\,R\,v$
  $B(B\varphi \rightarrow \varphi)$
- correctness of own beliefs:
  $B\neg B\varphi \rightarrow \neg B\varphi$
  for all w, there is a v such that $w\,R\,v$ and for all z if $v\,R\,z$ then $w\,R\,z$
  $BB\varphi \rightarrow B\varphi$
  density: for all w and v if $w\,R\,v$ then there is a z such that $w\,R\,z$ and $z\,R\,v$

Buridan-Burge Paradox I

Suppose that q is the statement that $\neg B_a q$.

Buridan-Burge Paradox I

Suppose that q is the statement that $\neg B_a q$. Now, either $B_a q$ or $\neg B_a q$.

Buridan-Burge Paradox I

Suppose that q is the statement that $\neg B_a q$. Now, either $B_a q$ or $\neg B_a q$.

1. Suppose $\neg B_a q$. Then by the 5 axiom ($\neg B_a \varphi \rightarrow B_a \neg B_a \varphi$), we have that $B_a \neg B_a q$. But since q is $\neg B_a q$, we have $B_a q$. Contradiction.

Buridan-Burge Paradox I

Suppose that q is the statement that $\neg B_a q$. Now, either $B_a q$ or $\neg B_a q$.

1. Suppose $\neg B_a q$. Then by the 5 axiom ($\neg B_a \varphi \rightarrow B_a \neg B_a \varphi$), we have that $B_a \neg B_a q$. But since q is $\neg B_a q$, we have $B_a q$. Contradiction.
2. Suppose $B_a q$. By by the 4 axiom ($B_a \varphi \rightarrow B_a B_a \varphi$), we have that $B_a B_a q$. By the D axioms ($B_a \varphi \rightarrow \neg B_a \neg \varphi$), we have that $\neg B_a \neg B_a q$. But since $\neg B_a q$ is q, we have $\neg B_a q$. Contradiction.

Tyler Burge (1984). *Epistemic paradox*. Journal of Philosophy, 81(1), pp. 5 - 29.

Buridan-Burge Paradox II

Of course, "q is the statement that $\neg B_a q$" is not a sentence of the modal logic of beliefs.

What we have shown is that $\neg B_a(q \leftrightarrow \neg B_a q)$ is a theorem of **KD45**.

This is a paradox only if it should be possible for an ideally rational agent to believe that $q \leftrightarrow \neg B_a q$.

Wolfgang Lenzen (1981). *Doxastic Logic and the Burge-Buridan-Paradox*. Philosophical Studies, 39(1), pp. 43 - 49.

Michael Caie (2012). *Belief and indeterminacy*. The Philosophical Review, 121(1), pp. 1 - 54.

T. Raleigh (2021). *A New Anti-Expertise Dilemma*. Synthese, 199, pp. 5551 - 5569.

Decision Instability

(DD) S Chooses $\phi \leftrightarrow \phi$ does not maximize utility for S

Decision Instability

(DD) S Chooses $\phi \leftrightarrow \phi$ does not maximize utility for S

Death in Damascus: Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him.

Decision Instability

(DD) S Chooses $\phi \leftrightarrow \phi$ does not maximize utility for S

Death in Damascus: Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...

Decision Instability

(DD) S Chooses $\phi \leftrightarrow \phi$ does not maximize utility for S

Death in Damascus: Death works from an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment on the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with Death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo...If he decides to go to Aleppo, he then has strong grounds for expecting that Aleppo is where Death already expects him to be, and hence it is rational for him to prefer staying in Damascus. Similarly, deciding to stay in Damascus would give him strong grounds for thinking that he ought to go to Aleppo. (Gibbard & Harper, 1978, p. 373)

Anti-Expert

$$(AE) \quad S \text{ Believes } p \leftrightarrow \neg p$$

Anti-Expert

$$(AE) \quad S \text{ Believes } p \leftrightarrow \neg p$$

"I'm a neurologist, and know there's a device that has been shown to induce the following state in people: they believe that their brains are in state S iff their brains are not in state S. I watch many trials with the device, and become extremely confident that it's extremely reliable.

Anti-Expert

$$(AE) \quad S \text{ Believes } p \leftrightarrow \neg p$$

"I'm a neurologist, and know there's a device that has been shown to induce the following state in people: they believe that their brains are in state S iff their brains are not in state S. I watch many trials with the device, and become extremely confident that it's extremely reliable. I'm also confident that my brain is not in state S.

Anti-Expert

$$(AE) \quad S \text{ Believes } p \leftrightarrow \neg p$$

"I'm a neurologist, and know there's a device that has been shown to induce the following state in people: they believe that their brains are in state S iff their brains are not in state S. I watch many trials with the device, and become extremely confident that it's extremely reliable. I'm also confident that my brain is not in state S.Then the device is placed on my head and switched on. My confidence that my brain is not in state S....well, it's not clear here what should happen here." (Christensen 2010, drawn from Conee 1982)

Propositional Quantifiers

While we naturally quantify over propositions in both ordinary and philosophical discussion of beliefs, the addition of propositional quantifiers is not given much attention in the literature.

Propositional Quantifiers

While we naturally quantify over propositions in both ordinary and philosophical discussion of beliefs, the addition of propositional quantifiers is not given much attention in the literature. Consider the following examples:

- "One believes that everything one believes is true": $B\forall p(Bp \to p)$
- "If no matter what p stands for, one believes that φ, then one believes that no matter what p stands for, φ": $\forall pB\varphi \to B\forall p\varphi$
- "There is a proposition that the agent takes to be consistent and to settle everything": $\exists q(Bq \wedge \forall p(B(q \to p) \vee B(q \to \neg p)))$

Immodest Beliefs

Immod: "One believes that everything one believes is true": $B\forall p(Bp \rightarrow p)$

- Even for idealized agents or idealized beliefs, as axiomatized by **KD45**, it seems that Immod should not be included in a logic of belief.

Immodest Beliefs

Immod: "One believes that everything one believes is true": $B\forall p(Bp \to p)$

- Even for idealized agents or idealized beliefs, as axiomatized by **KD45**, it seems that Immod should not be included in a logic of belief.

- `Immod` should be distinguished from "for every proposition p, one believes that if she believes that p then p": $\forall p(B(Bp \to p))$.

Immodest Beliefs

Immod: "One believes that everything one believes is true": $B\forall p(Bp \to p)$

- Even for idealized agents or idealized beliefs, as axiomatized by **KD45**, it seems that Immod should not be included in a logic of belief.

- `Immod` should be distinguished from "for every proposition p, one believes that if she believes that p then p": $\forall p(B(Bp \to p))$.

Consider an agent who has credences about a real number x randomly generated from the interval $[0, 1]$. For all measurable $X \subseteq [0, 1]$, the agent's credence that $x \in X$ is just the measure of X. Suppose that the agent outright believes precisely those propositions with credence 1. Then, for all $a \in [0, 1]$, the agent believes that $x \in [0, 1] \setminus \{a\}$ since $[0, 1] \setminus \{a\}$ is measure 1. However, the agent does not believe that for all $a \in [0, 1]$, $x \in [0, 1] \setminus \{a\}$ since $\bigcap_{a \in [0,1]}([0, 1] \setminus \{a\}) = \varnothing$, which is not measure 1.

Hence the agent in this situation does not believe that all her beliefs are true.

Yifeng Ding (2021). *On the Logic of Belief and Propositional Quantification*. Journal of Philosophical Logic, 50, pp. 1143 - 1198.

In any possible world semantics for **KD45**, $B\forall p(Bp \rightarrow p)$ is valid on any frame. So, any logic validating **KD45** must validate Immod. Algebraic semantics is needed for logics that do not validate Immod.

Yifeng Ding (2021). *On the Logic of Belief and Propositional Quantification*. Journal of Philosophical Logic, 50, pp. 1143 - 1198.

Also, see:

Jeremy Goodman (2020). *I'm mistaken*. manuscript.

Prior's Theorem

$$Q(\forall p(Qp \rightarrow \neg p)) \rightarrow (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$$

is a derivable using Universal Instantiation and propositional reasoning.

A. N. Prior. *On a family of paradoxes*. Notre Dame Journal of Formal Logic, 2(1), pgs. 16 - 32, 1961.

Prior's Theorem

T1. $C(UpCdpNp)\ C(dUpCdpNp)\ (NUpCdpNp)$ – from $CUpdpdq$ by substitution.
T2. $C(dUpCdpNp)\ C(UpCdpNp)\ (NUpCdpNp)$ – from $T1$ and $CCpCqrCqCpr$.
T3. $C(dUpCdpNp)\ (NUpCdpNp)$ – from $T2$ and $CCpCqNqCpNq$.
T4. $C(dUpCdpNp)(EpKdpp)$ – from $T3$ and equivalence of 'not-none' and 'some', i.e. of 'not-all-not' and 'some'.
T5. $C(dUpCdpNp)\ K(dUpCdpNp)\ (NUpCdpNp)$ – from $T3$ and $CCpqCpKpq$.
T6. $CK(dUpCdpNp)(NUpCdpNp)(EpKdpNp)$ – substitution in $CdqEpdp$.
T7. $C(dUpCdpNp)(EpKdpNp)$ – syllogistically from $T5$ and $T6$.
T8. $C(dUpCdpNp)\ K(EpKdpp)(EpKdpNp)$ – from $T4$, $T7$ and $CCpqCCprCpKqr$.

Prior's Theorem

1. $\forall p\,(Qp \to \neg p) \to (Q(\forall p(Qp \to \neg p)) \to \neg \forall p(Qp \to \neg p))$

   $(\forall p\,\varphi(p) \to \varphi[p/q])$

Prior's Theorem

1. $\forall p(Qp \to \neg p) \to (Q(\forall p(Qp \to \neg p)) \to \neg\forall p(Qp \to \neg p))$
   $(\forall p\varphi(p) \to \varphi[p/q])$
2. $Q(\forall p(Qp \to \neg p)) \to (\forall p(Qp \to \neg p) \to \neg\forall p(Qp \to \neg p))$
   $((a \to (b \to c)) \to (b \to (a \to c)))$

Prior's Theorem

1. $\forall p\,(Qp \to \neg p) \;\to (Q(\,\forall p(Qp \to \neg p)\,) \to \neg\,\forall p(Qp \to \neg p)\,)$
   $(\forall p \varphi(p) \to \varphi[p/q])$
2. $Q(\forall p(Qp \to \neg p)) \to (\,\forall p(Qp \to \neg p) \to \neg\forall p(Qp \to \neg p)\,)$
   $((\,a \to (b \to c)) \to (b \to (\,a \to c)))$
3. $Q(\forall p(Qp \to \neg p)) \to \neg\forall p(Qp \to \neg p)$
   $((a \to (b \to \neg b)) \to (a \to \neg b))$

Prior's Theorem

1. $\forall p\,(Qp \to \neg p) \;\to (Q(\,\forall p(Qp \to \neg p)\,) \to \neg\,\forall p(Qp \to \neg p)\,)$
   $(\forall p\varphi(p) \to \varphi[p/q])$
2. $Q(\forall p(Qp \to \neg p)) \;\to (\,\forall p(Qp \to \neg p) \;\to\; \neg\forall p(Qp \to \neg p)\,)$
   $((\,a \;\to (b \to c)) \to (b \to (\,a \;\to c)))$
3. $Q(\forall p(Qp \to \neg p)) \;\to$ $\neg\forall p(Qp \to \neg p)$
   $((a \to (b \to \neg b)) \to (a \to \neg b))$
4. $Q(\forall p(Qp \to \neg p)) \to$ $\exists p(Qp \wedge p)$
   ($\neg\forall p\varphi \leftrightarrow \exists p\neg\varphi$ and $\neg(a \to \neg b) \leftrightarrow (a \wedge b)$)

Prior's Theorem

1. $\forall p\,(Qp \to \neg p) \;\to (Q(\,\forall p(Qp \to \neg p)\,) \to \neg\,\forall p(Qp \to \neg p)\,)$
   $(\forall p\varphi(p) \to \varphi[p/q])$
2. $Q(\forall p(Qp \to \neg p)) \;\to (\,\forall p(Qp \to \neg p) \;\to\; \neg\forall p(Qp \to \neg p)\,)$
   $((\,a \;\to (b \to c)) \to (b \to (\,a \;\to c)))$
3. $Q(\forall p(Qp \to \neg p)) \to \neg\forall p(Qp \to \neg p)$
   $((a \to (b \to \neg b)) \to (a \to \neg b))$
4. $Q(\forall p(Qp \to \neg p)) \to\; \exists p(Qp \wedge p)$
   $(\;\neg\forall p\varphi \leftrightarrow \exists p\neg\varphi$ and $\neg(a \to \neg b) \leftrightarrow (a \wedge b))$
5. $Q(\forall p(Qp \to \neg p)) \to (\,Q(\forall p(Qp \to \neg p)) \wedge \neg\forall p(Qp \to \neg p)\,)$
   $((a \to b) \to (a \to (a \wedge b)))$

Prior's Theorem

1. $\forall p\,(Qp \to \neg p) \to (Q(\forall p(Qp \to \neg p)) \to \neg \forall p(Qp \to \neg p))$
   $(\forall p \varphi(p) \to \varphi[p/q])$
2. $Q(\forall p(Qp \to \neg p)) \to (\forall p(Qp \to \neg p) \to \neg \forall p(Qp \to \neg p))$
   $((a \to (b \to c)) \to (b \to (a \to c)))$
3. $Q(\forall p(Qp \to \neg p)) \to \neg \forall p(Qp \to \neg p)$
   $((a \to (b \to \neg b)) \to (a \to \neg b))$
4. $Q(\forall p(Qp \to \neg p)) \to \exists p(Qp \wedge p)$
   $(\neg \forall p \varphi \leftrightarrow \exists p \neg \varphi$ and $\neg(a \to \neg b) \leftrightarrow (a \wedge b))$
5. $Q(\forall p(Qp \to \neg p)) \to (Q(\forall p(Qp \to \neg p)) \wedge \neg \forall p(Qp \to \neg p))$
   $(a \to b) \to (a \to (a \wedge b))$
6. $Q(\forall p(Qp \to \neg p)) \to \exists p(Qp \wedge \neg p)$
   $((Q\varphi \wedge \neg \varphi) \to \exists p(Qp \wedge \neg p))$

Prior's Theorem

4. $Q(\forall p(Qp \to \neg p)) \to \exists p(Qp \wedge p)$
   ($\neg\forall p\varphi \leftrightarrow \exists p\neg\varphi$ and $\neg(a \to \neg b) \leftrightarrow (a \wedge b)$)
5. $Q(\forall p(Qp \to \neg p)) \to (Q(\forall p(Qp \to \neg p)) \wedge \neg\forall p(Qp \to \neg p))$
   $(a \to b) \to (a \to (a \wedge b))$
6. $Q(\forall p(Qp \to \neg p)) \to \exists p(Qp \wedge \neg p)$
   $((Q\varphi \wedge \neg\varphi) \to \exists p(Qp \wedge \neg p))$
7. $Q(\forall p(Qp \to \neg p)) \to (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$
   $(((a \to b) \wedge (a \to c)) \to (a \to (b \wedge c)))$

$$Q(\forall p(Qp \to \neg p)) \to (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$$

$$Q(\forall p(Qp \to \neg p)) \to (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$$

- $Q\varphi :=$ Ann believes that φ

  If Ann believes that everything that Ann believes is wrong, then Ann believes something true and Ann believes something wrong.

$$Q(\forall p(Qp \to \neg p)) \to (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$$

- $Q\varphi :=$ Ann believes that φ

  If Ann believes that everything that Ann believes is wrong, then Ann believes something true and Ann believes something wrong.

- $Q\varphi :=$ Ann says that φ

  If Ann says that everything that Ann says is wrong, then Ann says something true and Ann says something wrong.

$$Q(\forall p(Qp \to \neg p)) \to (\exists p(Qp \wedge p) \wedge \exists p(Qp \wedge \neg p))$$

- $Q\varphi :=$ Ann believes that φ

  If Ann believes that everything that Ann believes is wrong, then Ann believes something true and Ann believes something wrong.

- $Q\varphi :=$ Ann says that φ

  If Ann says that everything that Ann says is wrong, then Ann says something true and Ann says something wrong.

- $Q\varphi :=$ Ann wrote on the board at midnight that φ

  If Ann wrote on the board at midnight that everything that Ann wrote on the board at midnight is wrong, then Ann wrote a true thing on the board at midnight and Ann wrote a false thing on the board at midnight.

A. Bacon, J. Hawthorne and G. Uzquiano. *Higher-Order Free Logic and the Prior-Kaplan Paradox*. Forthcoming in *Williamson on Modality*.

A. Bacon and G. Uzquiano. *Some results on the limits of thought*. Journal of Philosophical Logic, 2018.

R. H. Thomason and D. Tucker. *Paradoxes of Intensionality*. Review of Symbolic Logic, 4, pgs. 394 - 411, 2011.

S5

K	$K(\varphi \to \psi) \to (K\varphi \to K\psi)$
T	$K\varphi \to \varphi$
4	$K\varphi \to KK\varphi$
5	$\neg K\varphi \to K\neg K\varphi$

The logic **S5** adds the above axiom schemes to an axiomatization of classical propositional logic with the rules Modus Ponens, Substitution of Equivalents, and Necessitation (from φ infer $K\varphi$).

S5 is sound and strongly complete with respect to all partition frames.

S4

K	$K(\varphi \to \psi) \to (K\varphi \to K\psi)$
T	$K\varphi \to \varphi$
4	$K\varphi \to KK\varphi$

The logic **S4** adds the above axiom schemes to an axiomatization of classical propositional logic with the rules Modus Ponens, Substitution of Equivalents, and Necessitation (from φ infer $K\varphi$).

S4 is sound and strongly complete with respect to all reflexive and transitive frames.

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

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: $\forall A(P(A) \rightarrow \Box A)$, where P is a predicate and $\Box$ is an operator.

- 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"

- 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: $\forall A(P(A) \rightarrow \Box A)$, 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. However, there is the question about 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.

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

$$\forall x \forall y((\mathsf{Sent}(x) \wedge \mathsf{Sent}(y)) \rightarrow (\Box \ulcorner x \dot{\rightarrow} y \urcorner \rightarrow (\Box x \rightarrow \Box y)))$$

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 **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 **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.

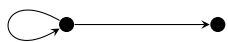
1. The following frame does not admit a valuation:

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

1. The following frame does not admit a valuation:

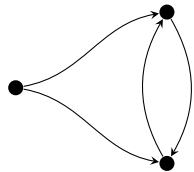
Use the fixed point: $A \leftrightarrow \neg\Box\ulcorner\Box\ulcorner A\urcorner\urcorner$

2. The following frame does not admit a valuation:



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

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:

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

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

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

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 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, $F \subseteq T \rightarrow F \subsetneq K$

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."

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.