Knowledge, Proof and the Knower

Walter Dean & Hidenori Kurokawa

(University of Warwick and CUNY) & CUNY

FEW 2009, CMU

The original Knower (Montague & Kaplan 1960)

- $\mathcal{L}_T \supseteq \mathcal{L}_a \cup \{K(x)\}$
- T an $\mathcal{L}_T$ theory s.t. $T \supseteq$ Q
- It follows that for every $\varphi(x) \in \text{Form}_{\mathcal{L}_T}$ there is γ s.t.

$$T \vdash \gamma \leftrightarrow \varphi(\ulcorner \gamma \urcorner)$$

- In particular, there is δ such that
  (FPm) $T \vdash \delta \leftrightarrow K(\ulcorner \neg\delta \urcorner)$.
- Additionally $K(x)$ is taken to satisfy
  (T) $K(\ulcorner \varphi \urcorner) \rightarrow \varphi$
  (U) $K(\ulcorner K(\ulcorner \varphi \urcorner) \rightarrow \varphi \urcorner)$
  (I) $(K(\ulcorner \varphi \urcorner) \wedge I(\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner)) \rightarrow K(\ulcorner \psi \urcorner)$
  where $T \vdash I(\ulcorner \varphi \urcorner, \ulcorner \psi \urcorner)$ iff $T \vdash \varphi \rightarrow \psi$

Original Knower (derivation)

Proposition

T is inconsistent.

1)	$T \vdash \delta \leftrightarrow \neg K(\ulcorner \neg\delta \urcorner)$	**FPm**
2)	$T \vdash K(\ulcorner \neg\delta \urcorner) \rightarrow \neg\delta$	**T**
3)	$T \vdash \delta \rightarrow \neg\delta$	1), 2)
4)	$T \vdash \neg\delta$	3)
5)	$T \vdash I(\ulcorner K(\ulcorner \neg\delta \urcorner) \rightarrow \neg\delta \urcorner, \ulcorner \neg\delta \urcorner)$	1) - 4)
6)	$T \vdash K(\ulcorner K(\ulcorner \neg\delta \urcorner) \rightarrow \neg\delta \urcorner)$	**U**
7)	$T \vdash [K(\ulcorner K(\ulcorner \neg\delta \urcorner) \rightarrow \neg\delta \urcorner) \wedge I(\ulcorner K(\ulcorner \neg\delta \urcorner) \rightarrow \neg\delta \urcorner, \ulcorner \neg\delta \urcorner))] \rightarrow K(\ulcorner \neg\delta \urcorner)$	**I**
8)	$T \vdash K(\ulcorner \neg\delta \urcorner)$	5), 6), 7)
9)	$T \vdash \delta$	1), 8)
10)	$T \vdash \bot$	4), 9)

Diagnosing the Paradox

theorist(s)	diagnosis	cure
Myhill [60]	informal provability is not expressible in $\mathcal{L}_a$	Forbid iterated modalities
Montague [63]	"syntactic treatment of modality"	treat modalities as sentential operators
Maitzen [98]	epistemic closure	Reject $\mathbf{I}$ -- i.e. $(K(\ulcorner\varphi\urcorner) \wedge I(\ulcorner\varphi\urcorner, \ulcorner\psi\urcorner)) \rightarrow K(\ulcorner\psi\urcorner)$
McGee [91], Horsten [02]	complicated	Reject $\mathbf{T}$ -- i.e. $K(\ulcorner\varphi\urcorner) \rightarrow \varphi$
Anderson [83], Cross [01] Égré [05]	process of elimination (?)	Reject $\mathbf{U}$ -- i.e. $K(\ulcorner K(\ulcorner\varphi\urcorner) \rightarrow \varphi\urcorner)$

A philosophical puzzle

T “knowledge entails truth”

$K(\ulcorner \varphi \urcorner) \to \varphi$

U “T is known”

$K(\ulcorner K(\ulcorner \varphi \urcorner) \to \varphi \urcorner)$

- Consensus view: retain **T** and **I**, reject **U**.
- **T** expresses (something like) a conceptual truth about knowledge.
- **U** appears to follow from **T** by reflecting on the meaning of “knows that.”
- So how can we accept **T** and reject **U**?

Plan

I) a simplified Knower
II) reconstruction in modal logic (S4)
III) reconstruction in **explicit** modal logic (QLP)
IV) isolation of a new principle (UBF) **needed to derive U from T**
V) arguments against UBF

A simplified Knower ($\approx$ Montague 1963)

- $\mathcal{L}_T \supseteq \mathcal{L}_a \cup \{K(x)\}$
- T an $\mathcal{L}_T$ theory s.t. $T \supseteq \mathrm{PA}$
- There is δ such that

  $(\mathbf{FP})\quad T \vdash \delta \leftrightarrow \neg K(\ulcorner \delta \urcorner)$
- For ease of reference

  $(\mathbf{FP1})\quad T \vdash \neg K(\ulcorner \delta \urcorner) \rightarrow \delta$

  $(\mathbf{FP2})\quad T \vdash K(\ulcorner \delta \urcorner) \rightarrow \neg\delta$
- We additionally suppose $K(x)$ satisfies

  $(\mathbf{T})\quad T \vdash K(\ulcorner \varphi \urcorner) \rightarrow \varphi$

  $(\mathbf{Int})$ If $T \vdash \varphi$, then $T \vdash K(\ulcorner \varphi \urcorner)$
- We will refer to **Int** as an **internalization principle**.
- NB: It resembles a traditional modal necessitation rule.

A simplified Knower (derivation)

Proposition

T is inconsistent.

1)	$T \vdash \neg K(\ulcorner \delta \urcorner) \rightarrow \delta$	**FP1**
2)	$T \vdash K(\ulcorner \delta \urcorner) \rightarrow \neg\delta$	**FP2**
3)	$T \vdash K(\ulcorner \delta \urcorner) \rightarrow \delta$	**T**
4)	$T \vdash \neg K(\ulcorner \delta \urcorner)$	2), 3)
5)	$T \vdash \delta$	1), 4)
6)	$T \vdash K(\ulcorner \delta \urcorner)$	**Int** 5)
7)	$T \vdash \bot$	4), 6)

Reasoning about knowledge as an operator

- The reasoning used in the Knower can be reconstructed if K is treated as a **sentential operator** rather than a predicate.
- This is the conventional approach of *epistemic logic* (Hintikka, Fagin et al.).
- Modal analogues to $\mathbf{T}$, $\mathbf{U}$, $\mathbf{I}$ and $\mathbf{Int}$:

  T $\Box F \rightarrow F$
  U $\Box(\Box F \rightarrow F)$
  K $\Box(F \rightarrow G) \rightarrow (\Box F \rightarrow \Box G)$
  Nec $\vdash F \quad \therefore \quad \vdash \Box F$

- But what is the status of $\mathbf{FP}$ on the operator interpretation?

$\Box\varphi \rightleftharpoons \varphi$ is knowable

$\Box\varphi \rightleftharpoons \varphi$ is provable

formal (i.e. in PA) | informal

GL (Smorynski, Boolos)

K: $\Box(F \to G) \to (\Box F \to \Box G)$
4: $\Box F \to \Box\Box F$
L: $\Box(\Box F \to F) \to \Box F$
Nec: $\vdash F \quad \therefore \ \vdash \Box F$
Reflection: **NO**
GL + **T** is inconsistent
Fixed points: **YES**
If P is modalized in Θ, then $\exists D$ s.t.

$$\mathrm{GL} \vdash D \leftrightarrow \Theta[D/P]$$

S4 (Gödel)

K: $\Box(F \to G) \to (\Box F \to \Box G)$
4: $\Box F \to \Box\Box F$
T: $\Box F \to F$
Nec: $\vdash F \quad \therefore \ \vdash \Box F$
Reflection: **YES**
T is an axiom
Fixed points: **NO**
$\mathrm{S4} \vdash \neg\Box(F \leftrightarrow \neg\Box F)$ for all F

The simplified Knower in S4

0)	$\Box(F \leftrightarrow \neg\Box F)$	$\vdash F \leftrightarrow \neg\Box F$	**T**
1)	"	$\vdash \neg\Box F \rightarrow F$	
2)	"	$\vdash \Box F \rightarrow \neg F$	
3)	"	$\vdash \Box F \rightarrow F$	**T**
4)	"	$\vdash \neg\Box F$	2), 3)
5)	"	$\vdash F$	1), 4)
6)	"	$\vdash \Box F$	**Nec** 5)
7)	"	$\vdash \bot$	4), 6)
8)		$\vdash \neg\Box(F \leftrightarrow \neg\Box F)$	0) - 7)

Proposition

For all F, S4 $\vdash \neg\Box(F \leftrightarrow \neg\Box F)$

Knowledge and justification

- Classically: F is known (by agent i) $\rightleftharpoons$
  i) F is true
  ii) i believes F
  iii) i is **justified** in believing F
  iv) $\ldots$
- Claim: the notion of knowledge relevant to the Knower is **knowledge in virtue of proof.**
- Why? Because derivability in T is intended to represent i's own (idealized) deductive capacities.
- This is implicit in the $\mathbf{I}$ and $\mathbf{Int}$.
- Question: What happens to the Knower if we try to represent justifications explicitly?

Making justification explicit

- Q: What are "**justifications**"?
- A: In the context of the Knower, it is reasonable to identify them with ***mathematical proofs***.
- t justifies $\varphi \rightleftharpoons \mathrm{Proof}_T(\ulcorner t \urcorner, \ulcorner \varphi \urcorner)$.
- **Explicit modalities:** $t : F \rightleftharpoons$ "t justifies F"
- Arithmetic interpretation: $(t : F)^* = \mathrm{Proof}_T(\ulcorner t^* \urcorner, \ulcorner F^* \urcorner)$
- For $T = \mathrm{PA}$, this interpretation leads to the **Logic of Proofs** [LP] (Artemov [01]).
- We'll work in a quantified extension of this system know as the **Quantified Logic of Proof** [QLP] (Fitting [04],[05]).

QLP (language)

- The class of QLP proof terms $Term_{QLP}$ is given by

$$t := x_i \mid a_i(x_{k_1}, \ldots, x_{k_n}) \mid\ !t \mid t_1 \cdot t_2 \mid t_1 + t_2 \mid \langle t \forall x \rangle$$

  - $x_1, x_2, x_3, \ldots$ are *proof variables*
  - $a_1(\vec{x}), a_2(\vec{x}), a_3(\vec{x}), \ldots$ are *primitive proof terms*
  - $!$, $\cdot$, $+$ and $\langle \cdot \forall \cdot \rangle$ are *proof operations*
- The class $Form_{QLP}$ of QLP formulas is given by

$$\varphi := P_i \mid F \wedge G \mid F \vee G \mid F \rightarrow G \mid \neg F \mid t : F \mid (\forall x)F \mid (\exists x)F$$

- Some characteristic formulas:
  - $a : ((F \wedge G) \rightarrow G)$ (a justifies $F \wedge G \rightarrow G$))
  - $!a : (a : ((F \wedge G) \rightarrow G))$ ($!a$ justifies $a : (F \wedge G \rightarrow G)$)
  - $x : (F \rightarrow G) \rightarrow (y : F \rightarrow x \cdot y : G)$
  - $b(x, y) : (x : (F \rightarrow G) \rightarrow (y : F \rightarrow x \cdot y : G))$
  - $(\forall x)(\forall y)[x : (F \rightarrow G) \rightarrow (y : F \rightarrow x \cdot y : G)]$

QLP (axioms)

LP1	all tautologies of classical propositional logic
LP2	$t : (F \to G) \to (s : F \to t \cdot s : G)$
LP3	$t : F \to F$
LP4	$t : F \to !t : t : F$
LP5	$t : F \to t + s : F$ and $s : F \to t + s : F$
QLP1	$(\forall x)F(x) \to F(t)$
QLP2	$(\forall x)(F \to G(x)) \to (F \to (\forall x)G(x))$
QLP3	$F(t) \to (\exists x)F(x)$
QLP4	$(\forall x)(F(x) \to G) \to ((\exists x)F(x) \to G)$
UBF	$(\forall x)t(x) : F(x) \to \langle t\forall x\rangle : (\forall x)F(x)$

- Usual definition of FV(F).
- Usual free-variable restrictions for QLP1-QLP4.

QLP (rules)

- A *primitive term specification* is a mapping $\mathcal{F}$ s.t. $\mathcal{P}(a(\vec{x}))$ is set of formulas $\mathcal{P}(a)$ such that if $F(\vec{x}) \in \mathcal{P}(a(\vec{x}))$, then $\mathrm{FV}(a) = \mathrm{FV}(\varphi)$. For all axioms $F(\vec{x})$, there is $a(\vec{x})$ s.t. $F(\vec{x}) \in \mathcal{P}(a(\vec{x}))$.
- Idea: if $F(\vec{x}) \in \mathcal{P}(a(\vec{x}))$, $a(\vec{x})$ serves a **name** for the axiom $F(\vec{x})$.
- QLP rules:
  - Modus Ponens
  - Axiom Internalization: if $F(\vec{x}) \in \mathcal{P}(a(\vec{x}))$, then $\vdash a(\vec{x}) : F(\vec{x})$
  - Universal Generalization: $\vdash F(x) \quad \therefore \quad \vdash (\forall x)F(x)$.
- A derived rule:
  - **JUG**: $\vdash t(x) : F(x) \quad \therefore \quad \vdash \langle t(x)\forall x\rangle : (\forall x)F(x)$

Necessitation vs. constructive internalization

- Traditional necessitation:
  - $\vdash F \quad \therefore \vdash \Box F$
  - idea: if F is derivable, then F is knowable/necessary/true
- Necessitation in QLP:
  - necessitation rule only for axioms

Proposition (Constructive Internalization Theorem)

If $\vdash F$, then there exists $t \in Term_{QLP}$ s.t. $\vdash t : F$.

- idea: if F is derivable in QLP, then there exists of F we can construct internally in QLP
- the axiom $t : (F \to G) \to (s : F \to t \cdot s : G)$ is used to internalize MP step
- UBF (or JUG) is used to internalize UG steps

Reconstructing the Knower in QLP (1)

Theorem (Realization (Artemov))

If S4 $\vdash F$, then there is an r s.t. LP $\vdash (F)^r$

where $(\cdot)^r$ uniformly replaces $\Box$s with terms $t \in Term_{LP}$.

Theorem (Embedding (Fitting))

If S4 $\vdash F$, then QLP $\vdash (F)^{\exists}$

where $(\Box F)^{\exists} = (\exists x)x : F^{\exists}$.

- So we should expect QLP to be incompatible with FPs.
- S4 $\vdash \neg\Box(F \leftrightarrow \neg\Box F) \Longrightarrow$

  QLP $\vdash \neg(\exists y)y : [F \leftrightarrow \neg(\exists x)x : F]$

Reconstructing the Knower in QLP (2)

0)	$y : (F \leftrightarrow \neg(\exists x)x : F)$	$\vdash F \leftrightarrow \neg(\exists x)x : F$	LP3
1)	"	$\vdash \neg(\exists x)x : F \rightarrow F$	
2)	"	$\vdash (\exists x)x : F \rightarrow \neg F$	
3)	"	$\vdash (\exists x)x : F \rightarrow F$	derivable in QLP
4)	"	$\vdash \neg(\exists x)x : F$	2), 3)
5)	"	$\vdash F$	1), 4)
6)	"	$\vdash t(y) : F$	for some $t(y)$ (via CIT)
6')	"	$\vdash (\exists x)x : F$	QLP3
7)	"	$\vdash \perp$	4), 6')
8)		$\vdash \neg y : (F \leftrightarrow \neg(\exists x)x : F)$	0) - 7)
9)		$\vdash (\forall y)\neg y : [F \leftrightarrow \neg(\exists x)x : F]$	UG
10)		$\vdash \neg(\exists y)y : [F \leftrightarrow \neg(\exists x)x : F]$	

Proposition

For all F, QLP $\vdash \neg(\exists y)y : [F \leftrightarrow \neg(\exists x)x : F]$

Reconstructing the Knower in QLP (3)

0) $y : (F \leftrightarrow \neg(\exists x)x : F) \vdash F \leftrightarrow \neg(\exists x)x : F$

1) " $\vdash \neg(\exists x)x : F \rightarrow F$

2) " $\vdash (\exists x)x : F \rightarrow \neg F$

3) " $\boldsymbol{\vdash (\exists x)x : F \rightarrow F}$ **derivable in QLP**

4) " $\vdash \neg(\exists x)x : F$

5) " $\vdash F$

6) " $\boldsymbol{\vdash t(y) : F}$ **for some $\boldsymbol{t(y)}$ (via CIT)**

6') " $\vdash (\exists x)x : F$

7) " $\vdash \bot$

8) $\vdash \neg y : (F \leftrightarrow \neg(\exists x)x : F)$

9) $\vdash (\forall y)\neg y : [F \leftrightarrow \neg(\exists x)x : F]$

10) $\vdash \neg(\exists y)y : [F \leftrightarrow \neg(\exists x)x : F]$

Reconstructing the Knower in QLP (4)

1) $\vdash x : F \rightarrow F$ LP3 (**explicit reflection**)
2) $\vdash (\forall x)(x : F \rightarrow F)$ UG
3) $\vdash (\forall x)(x : F \rightarrow F) \rightarrow ((\exists x)x : F \rightarrow F)$ QLP4
4) $\vdash (\exists x)(x : F \rightarrow F)$

---

1) $\vdash x : F \rightarrow F$ LP3
2) $\vdash r(x) : (x : F \rightarrow F)$ axiom nec.
3) $\vdash (\forall x)r(x) : (x : F \rightarrow F)$ UG
4) $\vdash (\forall x)r(x) : (x : F \rightarrow F) \rightarrow \langle r(x)\forall x\rangle : (\forall x)(x : F \rightarrow F)$ UBF
5) $\vdash \langle r(x)\forall x\rangle : (\forall x)(x : F \rightarrow F)$ 3),4)[1]
6) $\vdash q : (\forall x)(x : F \rightarrow F) \rightarrow ((\exists x)x : F \rightarrow F)$ axiom nec.
7) $\vdash q \cdot \langle r(x)\forall x\rangle : ((\exists x)x : F \rightarrow F)$ LP2 5), 6)

[1] We can get 5) from 2) via JUG:
$\vdash r(x) : (x : F \rightarrow F) \quad \therefore \quad \vdash \langle r(x)\forall x\rangle : (\forall x)(x : F \rightarrow F)$.

Reconstructing the Knower in QLP (6)

- We need to find $t(y)$ s.t. $y : (F \leftrightarrow \neg(\exists x)x : F) \vdash t(y) : F$.
- From above: $\vdash q \cdot \langle r(x)\forall x\rangle : ((\exists x)x : F \rightarrow F)$.
- We can take $t(y) \equiv (a_1 \cdot y) \cdot ((b \cdot (q \cdot \langle r(x)\forall x\rangle)) \cdot (a_2 \cdot y))$ (a_1, a_2 and b are constants for tautologies).
- Parallels:

  $$\begin{array}{llcll} \mathbf{T}_q & (\exists x)x : F \rightarrow F & \approx & \Box F \rightarrow F & \text{(i.e. } \mathbf{T}\text{)} \\ \mathbf{U}_q & (\exists y)y : ((\exists x)x : F \rightarrow F) & \approx & \Box(\Box F \rightarrow F) & \text{(i.e. } \mathbf{U}\text{)} \end{array}$$

- In the arithmetic and modal settings, **U** and **Int**/**Nec** are primitive.
- In QLP, $\mathbf{U}_q$ must be derived by constructive internalization.
- This appears to **require UBF** (or JUG) . . .

UBF, $\mathbf{U}_q$ and self-reference

- $\mathbf{U}_q \quad (\exists y)y : ((\exists x)x : F \to F) \quad \approx \quad \Box(\Box F \to F) \quad$ (i.e. **U**)
- QLP$^-$:= QLP - UBF, $\mathcal{L}_{QLP^-} = \mathcal{L}_{QLP} - \langle \cdot\forall\cdot \rangle$
- Using *Fitting semantics* for QLP we can show:

Proposition

If QLP $\nvdash F$, then QLP$^-$ $\nvdash (\exists y)y : ((\exists x)x : F \to F)$

Proposition

For any propositional letter P,

$$\text{QLP}^- + (\exists y)y : (P \leftrightarrow \neg(\exists x)x : P) \text{ is consistent.}$$

The arithmetic case against UBF and JUG

(1) $\vdash r(x) : (x : \bot \to \bot)$

(2) $\vdash (\forall x)r(x) : (x : \bot \to \bot)$

(3) $\vdash (\forall x)r(x) : (x : \bot \to \bot) \to \langle r(x)\forall x\rangle : (\forall x)(x : \bot \to \bot)$ UBF

(4) $\vdash \langle r(x)\forall x\rangle : (\forall x)(x : \bot \to \bot)$ 2), 3) via UBF or 1) via JUG

- Arithmetic interpretation of 1-2):
  $1^*)$ $\forall x[\mathrm{Proof}(r^*(x), \ulcorner \mathrm{Proof}(\dot{x}, \ulcorner \bot \urcorner) \to \bot \urcorner]$
- "For every $n \in \mathbb{N}$, we can prove that n is not a proof of $\bot$."
- $1^*)$ is **true** in N (presuming Z is **consistent**).
- Arithmetic interpreation of 3):

  $4^*)$ $\mathrm{Proof}(\langle r(x)\forall x\rangle^*, \ulcorner \forall x[\mathrm{Proof}(x, \ulcorner \bot \urcorner) \to \bot] \urcorner)$
- "The number $\langle r(x)\forall x\rangle^*$ **is a proof that Z is consistent.**"
- $4^*)$ is **false** in N because of Gödel's Second Incompleteness Theorem.

The conceptual case against UBF

1) $x : P \to P$
2) $r(x) : (x : P \to P)$
3) $(\forall x) r(x) : (x : P \to P)$
4) $\langle r(x) \forall x \rangle : (\forall x)(x : P \to P)$

- Claim: 1)-3) can all be true and 4) false.
- $P = \texttt{Substance } s \texttt{ contains cyanide.}$
- Proof variables denote chemical tests in a domain $\mathcal{D}_1$.
- It could be that all $t \in \mathcal{D}_1$ are **truth entailing** – i.e. such that $t : P \to P$ holds – and that we can prove this for each test we encounter.
- But there might be a larger domain $\mathcal{D}_2 \supsetneq \mathcal{D}_1$ of non-truth entailing tests (i.e. “non-standard” ones).
- If we can't **prove** that our quantifiers range over $\mathcal{D}_1$ and not over $\mathcal{D}_2$, then 4) should be false.

The provenance of UBF

- The original Barcan Formula:

(BF) $\forall x \Box \varphi(x) \to \Box \forall x \varphi(x)$

- If we take $\Box\varphi(x)$ as $\exists y \mathrm{Proof}(y, \ulcorner \varphi(x) \urcorner)$, BF corresponds to

(ω-rule) $\forall x \exists y \mathrm{Proof}(y, \ulcorner \varphi(\dot{x}) \urcorner) \to (\exists y)\mathrm{Proof}(y, \ulcorner \forall x \varphi(x) \urcorner)$

- If we take $\varphi(x) = \neg\mathrm{Proof}(x, \ulcorner \bot \urcorner)$, then ω-rule is **invalid** in N.
- Direct QLP analogue of ω-rule

(UBF$'$) $(\forall x)(\exists y) y : F(x) \to (\exists y) y : (\forall x) F(x)$

- UBF$'$ is **not** provable in QLP. UBF has a stronger antecedent:

(UBF) $(\forall x) t : F(x) \to \langle t \forall x \rangle : (\forall x) F(x)$

- But UBF already leads to **arithmetical unsoundness** ...

Arithmetical semantics for QLP$^-$ (1)

- Fix an injective primitive term specification $\mathcal{P}$.
- An arithmetic interpretation is a function $(\cdot)^* : \mathcal{L}_{QLP^-} \rightarrow \mathcal{L}_a$.
- Let Proof(x, y) be a PA proof predicate satisfying the usual GL conditions, $P = \{n \mid N \models \exists y < n(\mathrm{Proof}(\overline{n}, y))\}$, $p : \mathbb{N} \rightarrow \mathbb{N}$ a name for a p.r. function which enumerates P.
- $(\cdot)^*$ is defined inductively on proof terms as follows:
  - $(x_i)^* = p(x_i)$
  - if $F(\vec{x}) \in \mathcal{P}(a(\vec{x}))$, then $(a(\vec{x}))^* = a_F(\vec{x}^*) : \vec{\mathbb{N}} \rightarrow \mathbb{N}$
  - $(t \cdot s)^* = m(t^*, s^*), (t + s)^* = b(t^*, s^*), (!t)^* = c(t^*)$
- $(\cdot)^*$ is defined inductively on formulas as follows:
  - $P^* \in \mathrm{Sent}_{\mathcal{L}_a}$ (P atomic), $\bot^* = \bot$
  - $(F \rightarrow G)^* = F^* \rightarrow G^*, (\neg F)^* = \neg F^*$
  - $((\forall x)F)^* = (\forall x)F^*, ((\exists x)F)^* = (\exists x)F^*$
  - $\mathrm{Proof}(t^*, F(x_{k_1}, \ldots, x_{k_m})) =$ $\mathrm{Proof}(t^*, su(x^*_{k_m}, \overline{k}_m, \ldots, su(x^*_2, \overline{k}_2, su(x^*_1, \overline{k}_1, \ulcorner F(x^*_{k_1}, \ldots, x^*_{k_m})\urcorner)) \ldots$

Arithmetical semantics for QLP$^-$ (2)

- $(\cdot)^*$ is a $\mathcal{P}$-interpretation if $F \in \mathcal{P}(a) \Longrightarrow N \models (a : F)^*$.
- We say that a formula F is $\mathcal{P}$-valid if for all $\mathcal{P}$-interpretations $(\cdot)^*$, $N \models (F)^*$.
- If $\mathrm{FV}(F) = \{x_1, \ldots, x_n\}$, then $\mathrm{FV}(F)^* = \{x_1, \ldots, x_n\}$.
- $\mathcal{P}$-validity extends to open formulas as follows:

$$N \models (F(x_1, \ldots, x_n))^* \text{ iff for all } v,\ N \models_v (F(x_1, \ldots, x_n))^*$$

Proposition

There is a $\mathcal{P}$-interpretation $(\cdot)^*$ such that $N \models (\mathrm{QLP}^-)^*$.

<u>Proof</u>: E.g. $(x : F \rightarrow F)^* = \mathrm{Proof}(p(x), \ulcorner F^* \urcorner) \rightarrow F$. Suppose $N \models_v \mathrm{Proof}(p(x), \ulcorner F^* \urcorner)$. Then $N \models_v \mathrm{Proof}(p(x), \ulcorner F^* \urcorner)[\overline{v(x)}/x]$. Hence if $m = p(v(x))$, then $\mathrm{PA} \vdash \mathrm{Proof}(\overline{m}, \ulcorner F^* \urcorner)$. Thus $N \models F^*$.

Arithmetical soundness and UBF

- Let QLP$^\circ$ = all theorems of QLP not containing $\langle\cdot\forall\cdot\rangle$.
- NB: QLP$^-$ $\subsetneq$ QLP$^\circ$.
- The language of QLP$^\circ = \mathcal{L}_{QLP^-}$. So it makes sense to consider $(\mathrm{QLP}^\circ)^* = \{F^* \mid F \in \mathrm{QLP}^\circ\}$.
- As another diagnostic about UBF we have:

Proposition

For any $\mathcal{P}$-interpretation $(\cdot)^*$, PA $\cup$ $(\mathrm{QLP}^\circ)^*$ is inconsistent.

$PA \cup (QLP^\circ)^*$ is inconsistent.

Proof: Taking $F \equiv \bot$ above, note that

1) $QLP \vdash (\exists x)x : \bot \to \bot$
2) $QLP \vdash (\exists y)y : ((\exists x)x : \bot \to \bot)$

Note that since they do not contain $\langle \cdot \forall \cdot \rangle$, 1), 2) $\in QLP^\circ$.

3) $((\exists x)x : \bot \to \bot)^* = (\exists x)\mathrm{Proof}(p(x), \ulcorner \bot \urcorner) \to \bot$
4) $(\exists y)y : ((\exists x)x : \bot \to \bot)^* =$
$$(\exists y)[\mathrm{Proof}(p(y), \ulcorner (\exists x)(\mathrm{Proof}(p(x), \ulcorner \bot \urcorner) \to \bot) \urcorner]$$
5) $PA \vdash (\exists y)[\mathrm{Proof}(p(y), \ulcorner (\exists x)(\mathrm{Proof}(p(x), \ulcorner \bot \urcorner) \to \bot) \urcorner] \to$
$(\exists x)\mathrm{Proof}(p(x), \ulcorner \bot \urcorner)$ (Löb)
6) $PA \cup (QLP^\circ)^* \vdash (\exists x)\mathrm{Proof}(p(x), \ulcorner \bot \urcorner)$ 2), 4), 5)
7) $PA \cup (QLP^\circ)^* \vdash \bot$ 3), 6)

Morals about the Knower

- If we take $\Box$ to express knowability, we cannot drop **T**.
- We still get a paradox even if we drop **I** (or **K**) [Cross].
- Anderson, Érgé: Don't internalize **T** to yield **U**.
- **But what's wrong with U**?
- In QLP **U is not an axiom.**
- UBF or JUG is required to derive $\mathbf{U}_q$ from $\mathbf{T}_q$.
- We've seen four reasons to be suspicious of UBF / JUG:
  1) non-converstativeness
  2) consistency of $\mathrm{QLP}^- + (\exists y)y : (P \leftrightarrow \neg(\exists x)x : P)$
  3) inconsistency of $\mathrm{QLP}^\circ \cup \mathrm{PA}$
  4) the possibility that our proof quantifiers range over non-truth entailing justifications
- Upshot: The problem does lie with internalization, **but only of UG inferences.**

UBF and Gödel's "Lecture at Zilsel's"

- Gödel [38a] considers how to axiomatize the relation

$$z\mathrm{B}\varphi \rightleftharpoons z \text{ is a derivation of } \varphi$$

- Here B must express the "absolute" proof relation.
- In particular, Gödel appears to claim
  4) $\vdash a\mathrm{B}((\forall u)\neg u\mathrm{B}(0=1))$

  – i.e. a is a derivation of "no u is a derivation of $(0=1)$."
- His other axioms allow him to derive
  1) $\vdash \neg u\mathrm{B}(0=1)$
  2) $\vdash b\mathrm{B}(\neg u\mathrm{B}(0=1))$
  3) $\vdash (\forall u)b\mathrm{B}(\neg u\mathrm{B}(0=1))$ (?no UG rule is stated?)
- The step from 3) to 4) appears to require

(UBFg)

$(\forall u)z\mathrm{B}\varphi \rightarrow v\mathrm{B}((\forall u)\varphi)$ for some derivation v