On fixed points, diagonalization, and self-reference

Bernd Buldt

Department of Philosophy
Indiana U - Purdue U Fort Wayne (IPFW)
Fort Wayne, IN, USA
e-mail: buldtb@ipfw.edu

CL 16 – Hamburg – September 12, 2016

Section I: G1 & Fixed Points

G1 Proof, using the Gödel fixed point

Assumptions

(ADQ) $\vdash_{\mathcal{F}} \varphi \quad \Leftrightarrow \quad \vdash_{\mathcal{F}} \mathsf{Pr}_{\mathsf{F}}(\ulcorner \varphi \urcorner)$, for all $\varphi \in \mathcal{L}_{\mathcal{F}}$

(FPE) $\vdash_{\mathcal{F}} \gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner)$, for at least one $\gamma \in \mathcal{L}_{\mathcal{F}}$

Proof

$$\vdash_{\mathcal{F}} \gamma \overset{\text{ADQ}}{\Rightarrow} \vdash_{\mathcal{F}} \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner) \overset{\text{FPE}}{\Rightarrow} \vdash_{\mathcal{F}} \neg\gamma \Rightarrow \lightning \overset{\text{con}\,\mathcal{F}}{\Rightarrow} \nvdash_{\mathcal{F}} \gamma$$

$$\vdash_{\mathcal{F}} \neg\gamma \overset{\text{FPE}}{\Rightarrow} \vdash_{\mathcal{F}} \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner) \overset{\text{ADQ}}{\Rightarrow} \vdash_{\mathcal{F}} \gamma \Rightarrow \lightning \overset{\text{con}\,\mathcal{F}}{\Rightarrow} \nvdash_{\mathcal{F}} \neg\gamma$$

Fixed point derivation, Step 1: Substitution

- Fix a certain individual variable of your choice; say 'u.'
- Define a function *sub* that mirrors the substitution of the replacee variable 'u' for a replacer term 't,'
$$\varphi[\mathrm{u}]\tfrac{\mathrm{t}}{\mathrm{u}} \equiv \varphi(\mathrm{t}),$$
but in the realm of Gödel numbers. In short:
$$sub(x, y) := \begin{cases} gn(\varphi[\mathrm{u}]\tfrac{\bar{\mathrm{t}}}{\mathrm{u}}) & \text{if } x = gn(\varphi(\mathrm{u})) \text{ and } y = gn(\bar{\mathrm{t}}) \\ x & \text{otherwise.} \end{cases}$$
- Note that $sub(x, y)$ is primitive recursive and therefore represented by an expression $\varphi_{\mathrm{s}}(\mathrm{x}, \mathrm{y})$ in $\mathcal{F}$.

Fixed point derivation, Step 2: Definitions

- Define $\varphi(\mathsf{u}) :\equiv \forall \mathsf{x}\big[\neg \mathsf{Proof_F}(\mathsf{x}, \mathsf{sub}(\mathsf{u}, \mathsf{u}))\big]$.
- Define $p := gn(\varphi(\mathsf{u}))$.
- Substitute p for u in $\varphi(\mathsf{u})$, *viz.*,

$$\gamma :\equiv \varphi(\overline{\mathsf{p}}) \equiv \forall \mathrm{x}[\neg \mathrm{Proof_F}(\mathrm{x}, \mathrm{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))].$$

- Calculate

$$\begin{array}{rcll} sub(p,p) & = & sub\big(gn(\varphi(\mathsf{u})), p\big) & ;\ \text{def. } p \\ & = & gn\big(\varphi[\mathsf{u}]\frac{\overline{\mathsf{p}}}{\mathsf{u}}\big) & ;\ \text{def. } sub \\ & = & gn\big(\varphi(\overline{\mathsf{p}})\big) & ;\ \text{substitution} \\ & = & gn(\gamma) & ;\ \text{def. } \gamma \end{array}$$

Fixed point derivation, Step 3: Derivation

- Recall Step 2: $sub(p, p) = gn(\gamma)$.
- Reason inside $\mathcal{F}$.

$$\begin{array}{lll}
\vdash_{\mathcal{F}} \neg \mathsf{Pr_F}(\mathsf{x}) \leftrightarrow \neg \mathsf{Pr_F}(\mathsf{x}) & ; & \text{logic} \\
\vdash_{\mathcal{F}} \neg \mathsf{Pr_F}(\mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}})) \leftrightarrow \neg \mathsf{Pr_F}(\ulcorner \gamma \urcorner) & ; & \text{Step 2} \\
\vdash_{\mathcal{F}} \forall \mathrm{x} \left[\neg \mathrm{Proof_F}(\mathrm{x}, \mathrm{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}})) \right] \leftrightarrow \neg \mathsf{Pr_F}(\ulcorner \gamma \urcorner) & ; & \text{def. } \mathsf{Pr_F} \\
\vdash_{\mathcal{F}} \varphi(\overline{\mathsf{p}}) \leftrightarrow \neg \mathsf{Pr_F}(\ulcorner \gamma \urcorner) & ; & \text{def. } \varphi(\overline{\mathsf{p}}) \\
\vdash_{\mathcal{F}} \gamma \leftrightarrow \neg \mathsf{Pr_F}(\ulcorner \gamma \urcorner) & ; & \text{def. } \gamma
\end{array}$$

- Warning. We assumed $\vdash_{\mathcal{F}} \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}) = \ulcorner \gamma \urcorner$, which requires induction.

Theorem (Fixed Point Theorem, Diagonalization Lemma)

Assume $\mathcal{F}$ to allow for representation. For each expression φ with at least one variable free, there is a ψ such that,

$$\vdash_{\mathcal{F}} \psi \leftrightarrow \varphi_\psi$$

where φ_ψ can be either of the four forms:

$$\varphi(\ulcorner \psi \urcorner),\ \varphi(\ulcorner \neg\psi \urcorner),\ \neg\varphi(\ulcorner \psi \urcorner), \neg\varphi(\ulcorner \neg\psi \urcorner),$$

viz., instances of what we call a Henkin, Jeroslov, Gödel, or Rogers fixed point resp.

Proof.

Same as above (with minor modifications). □

Black self-referential magic?

- Two questions about fixed points such as
$$\vdash_{\mathcal{F}} \gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner).$$

1. How much "black magic" is required for their derivation?
   ... will be answered in Section II.

2. How much "self-reference" do they involve?
   ... will be answered in Section III.

Section II: Diagonalization

Black magic?

1^{st} Question

How much "black magic" is required for the derivation of fixed points such as

$$\vdash_{\mathcal{F}} \gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner)\,?$$

Answer

None.

Diagonalization

- Let $\mathcal{A} = \{a_{ij}\}_{i,j\in\omega}$ be a (countable) two-dimensional array:

$$\begin{array}{lccccc}
R_0: & a_{00} & a_{01} & \dots & a_{0n} & \dots \\
R_1: & a_{10} & a_{11} & \dots & a_{1n} & \dots \\
 & \vdots & \vdots & \ddots & \vdots & \\
R_n: & a_{n0} & a_{n1} & \dots & a_{nn} & \dots \\
 & \vdots & \vdots & & \vdots & \ddots
\end{array}$$

- Let f be a sequence transforming function,
$$f(R_n) = \{f(a_{ni})\}_{i\in\omega}.$$
- Apply f to the diagonal sequence D:
$$D' = f(D) := \langle f(a_{00}), f(a_{11}), f(a_{22}), \dots, f(a_{nn}), \dots\rangle.$$

Diagonalization: (Non-)Closure

- One of two things can happen to the anti-diagonal $D' = f(D)$:
  1. D' is identical to one of the rows, *viz.*, $f(D) = R_i \in \mathcal{A}$, for some i.
  2. D' is not identical to any of the rows, *viz.*, $f(D) \neq R_i \in \mathcal{A}$, for all i.
- If Case 1 applies, we call the set A closed under f, and f will have fixed points.
- If Case 2 applies, A is not closed under f, and we have Cantor's diagonal argument showing that a certain sequence is not in $\mathcal{A}$ (to "diagonalize out").

Diagonalization: Case 1 – Closure

- D' is identical to one of the rows, *viz.*, $f(D) = R_i \in \mathcal{A}$, for some i.

- The identity $D' = f(D) = R_i$ is element-wise identity:

$$
\begin{array}{cccccccc}
D' = \langle & f(a_{00}), & f(a_{11}), & \ldots, & f(a_{ii}), & \ldots, & f(a_{nn}), & \ldots \rangle \\
& \| & \| & & \| & & \| & \\
R_i = \langle & a_{i0}, & a_{i1}, & \ldots, & a_{ii}, & \ldots, & a_{in}, & \ldots \rangle
\end{array}
$$

- Closure under f (failure to "diagonalize out") implies fixed points $f(a_{ii}) = a_{ii}$.

Diagonalization: Case 1 – Closure

$$\begin{array}{lllllll} R_0: & a_{00} & a_{01} & \dots & a_{0n} & \dots \\ R_1: & a_{10} & a_{11} & \dots & a_{1n} & \dots \\ & \vdots & \vdots & \ddots & \vdots & \\ R_n: & a_{n0} & a_{n1} & \dots & a_{nn} & \dots \\ & \vdots & \vdots & & \vdots & \ddots \end{array} \Rightarrow \begin{array}{lllllll} R_0: & fa_{00} & a_{01} & \dots & a_{0n} & \dots \\ R_1: & a_{10} & fa_{11} & \dots & a_{1n} & \dots \\ & \vdots & \vdots & \ddots & \vdots & \\ R_n: & a_{n0} & a_{n1} & \dots & fa_{nn} & \dots \\ & \vdots & \vdots & & \vdots & \ddots \end{array}$$

$$\Rightarrow \begin{array}{lllllllll} R_0: & a_{00} & a_{01} & \dots & a_{0i} & \dots & a_{0n} & \dots \\ R_1: & a_{10} & a_{11} & \dots & a_{1i} & \dots & a_{1n} & \dots \\ & \vdots & \vdots & \ddots & \vdots & & \vdots & \\ f(D) = R_i: & \underline{fa_{00}}_{\,a_{i0}} & \underline{fa_{11}}_{\,a_{i1}} & \dots & \underline{fa_{ii}}_{\,a_{ii}} & \dots & \underline{fa_{nn}}_{\,a_{in}} & \dots \\ & \vdots & \vdots & & \vdots & \ddots & \vdots & \\ R_n: & a_{n0} & a_{n1} & \dots & a_{ni} & \dots & a_{nn} & \dots \end{array}$$

Diagonalization: Closure & Gödel fixed point

- Can we understand $\gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner)$ to be an instance of $f(a_{ii}) = a_{ii}$ for some f and some array $\mathcal{A} = \{a_{ij}\}_{i,j \in \omega}$?
- Yes.

Diagonalization: Closure & Gödel fixed points

- Step 1: Choose all first-order expressions with the free variable ‘u:’
$$A = \{\varphi_0(\mathrm{u}), \varphi_1(\mathrm{u}), \varphi_2(\mathrm{u}), \ldots\}.$$
- Step 2: Form the set of all of their Gödel numbers:
$$B = \{\ulcorner\varphi_0(\mathrm{u})\urcorner, \ulcorner\varphi_1(\mathrm{u})\urcorner, \ulcorner\varphi_2(\mathrm{u})\urcorner, \ldots\}.$$
- Step 3: Systematically plug all members of B into the free variable slots of all members of A; call this set C. We write ‘φ_{ab}’ instead of ‘$\varphi_a(\ulcorner\varphi_b\urcorner)$.’

Diagonalization: Gödel fixed points – 1^{st} diagonalization

- Lay out the elements of C in such a way that A determines the rows and B the columns which gives us::

$$
\begin{array}{cccccc}
 & \ulcorner\varphi_0\urcorner & \ulcorner\varphi_1\urcorner & & \ulcorner\varphi_n\urcorner & \\
\varphi_0 & \varphi_{00} & \varphi_{01} & \cdots & \varphi_{0n} & \cdots \\
\varphi_1 & \varphi_{10} & \varphi_{11} & \cdots & \varphi_{1n} & \cdots \\
 & \vdots & \vdots & \ddots & \vdots & \\
\varphi_n & \varphi_{n0} & \varphi_{n1} & \cdots & \varphi_{nn} & \cdots \\
 & \vdots & \vdots & & \vdots & \ddots
\end{array}
$$

- Note that the diagonal sequence $\{\varphi_{xx}\}_{x\in\omega}$ corresponds to the substitution function $sub(x, x)$ we used above.

Diagonalization: Gödel fixed points – 2^{nd} diagonalization

1. Observe that the provability predicate $\neg\mathsf{Pr_F}(\mathsf{u})$ is itself part of the first set we started out with: $A = \{\varphi_0, \varphi_1, \varphi_2, \ldots\}$; i. e., $\exists i$ s. t.: $\varphi_i \equiv \neg\mathrm{Pr_F}(\mathrm{u})$.
2. Apply the transformation $f : \varphi_{\mathrm{ab}} \mapsto \neg\mathrm{Pr_F}(\varphi_{\mathrm{ab}})$.
3. Because of (1), f maps C onto C, C will be closed under f, and each image $\neg\mathsf{Pr_F}(\varphi_{ab})$ must be a φ_{in}, for some n.
4. Hence, $f(D)$ has a fixed point φ_{ii}, which corresponds to the expression $\gamma \equiv \varphi(\overline{\mathsf{p}})$ we used above.

Diagonalization: Gödel fixed points without "black magic"

- Derivable fixed points in systems of arithmetic $\mathcal{F}_{Ar}$, e. g.,

  $$\gamma \leftrightarrow \neg \mathrm{Pr}_{\mathrm{F}}(\ulcorner \gamma \urcorner),$$

  are a result of the fact that set of expressions, such as A, are closed under certain transformations f.
  - $sub(x, x)$ corresponds to $\{\varphi_{xx}\}_{x \in \omega}$.
  - $\gamma \equiv \varphi(\overline{\mathrm{p}})$ corresponds to φ_{ii}.
  - Outcomes can be modelled in $\mathcal{F}_{Ar}$.
- The procedure ("double diagonalization") is entirely syntactic is completely mundane, no magic anywhere.

Section III: Self-Reference

Black magic?

2^{nd} Question

How much "self-reference" is required for the derivation of fixed points such as:

$$\vdash_{\mathcal{F}} \gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner)\,?$$

Answer

None.

Self-Reference: Rendered moot by diagonalization

- Previous section: Fixed points such as:

  $$\gamma \leftrightarrow \neg \mathsf{Pr}_{\mathsf{F}}(\ulcorner \gamma \urcorner),$$

  result from certain closure properties.

- The crucial steps,
  - $sub(x, x)$ or $\{\varphi_{xx}\}_{x \in \omega}$.
  - $\gamma \equiv \varphi(\overline{\mathsf{p}})$ or φ_{ii}.

  are entirely syntactic operations, which neither employ nor presuppose any concept of self-reference.

Self-Reference: Digging deeper

- Does $\psi \leftrightarrow \varphi(\psi)$ mean that ψ says it has property φ?
  - Does $\gamma \leftrightarrow \neg\mathsf{Pr_F}(\ulcorner\gamma\urcorner)$ mean that γ expresses some property it itself has, namely, the property "$\neg\mathsf{Pr_F}(\mathsf{u})$" (unprovability)?
  - If so, does it mean that γ states its own unprovability?
- Preliminaries: What self-reference cannot be.
  - Self-reference cannot mean γ is somehow a proper part of itself; this would violate the mereological definition of proper parthood, $PPxy := Pxy \wedge x \neq y$.
  - Self-reference hence presupposes a more abstract semantical relation than self-inclusion is.

Self-Reference: 'Propertual' self-reference

- Expression $\varphi(\mathrm{u})$ defines, in some structure $\mathfrak{A}$, property P if:
  1. Definition: $\{x : P(x)\}$ iff $\{x : \mathfrak{A} \models \varphi(\#\mathrm{x})\}$.

  Then $\varphi(\mathrm{u})$ has property P itself if:

  2. Self-Reference: $\mathfrak{A} \models \varphi(\#\varphi(\mathrm{u}))$.

- Application to $\neg\mathrm{Pr}_{\mathsf{F}}(\mathrm{u})$
  - $\mathfrak{N} \models \neg\mathrm{Pr}_{\mathsf{F}}(\ulcorner\neg\mathrm{Pr}_{\mathsf{F}}(\mathrm{u})\urcorner)$, because $\not\vdash_{\mathcal{F}} \neg\mathrm{Pr}_{\mathsf{F}}(\mathrm{u})$
  - Given suitable circumstances, 'propertual' self-reference may occur.
  - Mute point: no mention of $\gamma \leftrightarrow \neg\mathrm{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner)$.

Self-Reference: Propertual self-reference

- Problem. What conditions would elevate ψ in $\psi \leftrightarrow \varphi_\psi$ from being merely truth-functionally equivalent to actually being self-referential the same way φ_ψ is?
- All known attempts to identify such conditions can be considered to have failed, mostly because we do not yet have a good theory of self-reference.
  (see Halbach and Visser 2015)

Self-Reference: Improper self-reference

Direct objectual self-reference: $\varphi(\#\varphi)$; eg, *viz.*, $\varphi^\frown|\varphi|$, or $\varphi(\ulcorner\varphi\urcorner)$.

- Does γ in $\gamma \leftrightarrow \neg\mathsf{Pr}_\mathsf{F}(\ulcorner\gamma\urcorner)$ contain its own name?
- Recall that γ is shorthand for $\forall \mathsf{x}[\neg\mathsf{Proof}_\mathsf{F}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))]$, with $p = gn(\neg\mathsf{Pr}_\mathsf{F}(\mathsf{sub}(\mathsf{u}, \mathsf{u}))$.
- Thus, no.
- However, since $sub(\overline{\mathsf{p}}, \overline{\mathsf{p}}) = gn(\gamma)$, we know that γ would be self-referential if criteria would be more lax.

Self-Reference: Improper self-reference

Indirect objectual self-reference: $\varphi(\#\#\varphi)$; eg, $\varphi(\mathrm{t})$, with $t = \#\#\varphi(\mathrm{t})$

- Does γ in $\gamma \leftrightarrow \neg\mathrm{Pr_F}(\ulcorner\gamma\urcorner)$ contain its own indirect name?
- Since $sub(\overline{\mathrm{p}}, \overline{\mathrm{p}}) = gn(\gamma)$, the expression γ, which is $\forall \mathrm{x}[\neg\mathrm{Proof_F}(\mathrm{x}, \mathrm{sub}(\overline{\mathrm{p}}, \overline{\mathrm{p}}))]$, contains an indirect name of itself.
- Some (eg, Heck 2007) are perfectly happy to embrace the last point and call the Gödel sentence γ self-referential in the above sense and have it say "I'm not provable."

Self-Reference: Improper self-reference

- γ does not say "I" but refers to itself indirectly via a functional expression
- γ is true *iff* γ is not formally provable. By itself, this is a raw datum about γ's model theoretic evaluation and the resulting truth value. As such, it is just another equivalence that implies nothing about meaning or self-reference.
- Semantic stance like intentional stance; useful but not justified
- We practice semantic hunches, but gut feelings are a poor substitute for an actual theory.

Self-Reference: Summary

- Diagonalization produces fixed points.
- Fixed points do not establish self-reference.
- Self-reference we find is not proper internal self-reference, but our external attribution.

Thank You!