Indiana University - Purdue University Fort Wayne
Opus: Research & Creativity at IPFW

Philosophy Faculty Publications

Department of Philosophy

Summer 2016

On Fixed Points, Diagonalization, and Self-Reference

Bernd Buldt
Indiana University - Purdue University Fort Wayne, buldtb@ipfw.edu

Opus Citation

Bernd Buldt (2016). On Fixed Points, Diagonalization, and Self-Reference. *Von Rang und Namen. Essays in Honour of Wolfgang Spohn, ed. by Wolfgang Freitag et al.*. 47-63. Münster: Mentis.
http://opus.ipfw.edu/philos_facpubs/326

On fixed points, diagonalization, and self-reference.[*]

Bernd Buldt

December 28, 2015

Abstract

We clarify the respective roles fixed points, diagonalization, and self-reference play in proofs of Gödel's first incompleteness theorem.

The character of a *Festschrift*, I understand, allows for personal recollections of the honoree, so indulge me when I start out with a few of those . . .

It must have been 1998–99—Volker Halbach had joined our ranks in 1997 and we had started our weekly logic colloquium with Ulf Friedrichsdorf, a colleague in mathematics—when I stepped into Wolfgang's office and saw a copy of Boolos' *The Logic of Provability* on his desk. Me: "Oh, what's this about?" Wolfgang: "I want to check out what you guys are doing," followed by an impromptu sketch of his own completeness proof for modal logic he had devised while still being a graduate student but never published. I do not know whether he liked the book; I never asked. But there was another area where I remember Wolfgang ready to defend the importance of diagonalization, namely, in the philosophy of language and here in particular Kaplan's explication of intensions as diagonals of characters.

Despite my sustained interest in formal methods, I have never been able to ignore what makes philosophy a discipline in the humanities. That was one of the reasons to write my *Habilitation* on a historical topic. It speaks to Wolfgang's greatness as a person that he tolerated my heresy. Later, however, he would force me to speak on Bayesianism and the catch-all problem at my *Habilitationskolloquium* to find out whether all was lost; he probably concluded it was.

So, when I was sifting through potential topics for my contribution, anything even remotely historical or in a humanistic vein was out of the question and diagonalization came out on top as something we share an interest in. I developed the line of thought presented below in response to a blog post by Richard Zach and an unpublished draft by Richard Heck, included it to a lecture once, but would otherwise not have taken the time to write it up. I am therefore grateful to the editors for giving me an incentive to do so. True to its character as an incidental piece, I suppress almost all scholarly references; they can be found in Buldt 2014.

[*] Uncorrected draft. Final version in Wolfgang Freitag et al., *Von Rang und Namen. Essays in Honour of Wolfgang Spohn*, Münster: mentis (2016), 47–63 (page numbers tentative).

1 Introduction

Despite the fact that generations of researchers have vetted Gödel's result, there are some still who harbor a lurking suspicion that his incompleteness proof flirts with paradox or claim it to be one. Graham Priest, for example, built the entire cottage industry of paraconsistent logic and dialetheism on his initial analysis that a formally undecidable Gödel sentence is both true and false. But whether we look to Priest or others, the prime suspect to have facilitated the crime of paradox is always the allegedly self-referential Gödel sentence stating "I'm not provable." Some logicians tried to remove the aura of paradox by eliminating self-reference or by making its mechanisms more transparent. Many, however, continued to use "self-reference" in highly visible places like book titles despite the fact that we do not seem to have a good grasp of how to make the intuitive idea of self-reference sufficiently precise in formal contexts. Others take offense at diagonalization, intimating that it amounts to "black magic" (Soare) or at least is "intuitively unclear" (Kotlarski). Oftentimes the issue is further compounded by a somewhat loose language, which leaves unclear what "Gödel's incompleteness proof" really refers to.

What I hope to achieve in this paper, then, is to alleviate the situation just sketched by clarifying the respective roles fixed points, diagonalization, and self-reference play in Gödel's proof. This, I hope, will also refute the allegation that Gödel skates on the thin ice of paradox or that diagonalization is unintuitive. The exposition is organized around four claims: (*i*) the importance of fixed points; (*ii*) diagonalization as the technique of choice for fixed point construction; (*iii*) self-reference as a means for fixed point construction; (*iv*) distinctness of all three (e. g., diagonalization must not result in fixed points or fixed points not be self-referential).

2 Disclaimer

There is a robust consensus among logicians that Gödel's first theorem, properly conceived, is not a result in proof theory (in whose context it was first formulated) but a result in computability or recursion theory. There, its informal version reads:

> For every effective method that generates only true sentences of arithmetic we can effectively determine a true sentence that this method cannot generate.

This statement, translated into the language of recursion theory, becomes,

> The set of (Gödel numbers of) all true arithmetical sentences is productive.

Let a Gödel numbering be given and *TA* be True Arithmetic, here understood to be the set of Gödel numbers of all true arithmetical sentences. Productivity means that there is a total computable function f such that whenever i is the index of a computable subset W_i of *TA*: $W_i \subseteq TA$, then $f(i)$ lies in W_i's complement relative to *TA*: $f(i) \in TA \setminus W_i$. In other words, $f(i)$ is the Gödel number of a true arithmetical sentence (since it is in *TA*) but not generated by the ith method (since it is not in W_i).

Thus, any proof theoretic version of Gödel's first theorem becomes an instance of this more general result for one particular vehicle of calculation, *viz.*, the formal system encoded by the index i.

To my knowledge no one has yet argued that $f(i)$ is the Gödel number of both a true and a false sentence. And while this may raise concerns about the legitimacy of claims put forward by Priest and others, we also understand that if the mountain will not come to Muhammad, Muhammad will go to the mountain. This is why in the sections to follow we meet the infidels where they dwell, among the proof theoretic versions of Gödel's first theorem.

3 Preliminaries

This section can be safely skipped by everyone, who does not need a reminder of selected basic facts, for all we do in this section is fix some terminology and notation.

3.1 Formal Systems

Conceive of a formal system $\mathcal{F}$ as a system of fully formalized axiomatic reasoning. To be more specific, we identify a formal system $\mathcal{F}$ with a triple $\langle \mathcal{L}, \Sigma, R \rangle$, where $\mathcal{L}$ is a formal language, $\Sigma \subseteq \mathcal{L}$ is a set of axioms, possibly empty, and R is a set of (logic) rules defined over $\mathcal{L}$. We require all three components to be effectively given, *viz.*, language and rules are effectively decidable (i. e., recursive) and the axioms can be effectively listed (i. e., recursively enumerable, which, by Craig's well-known theorem, means Σ can chosen to be primitive recursive).

We call an expression φ derivable or formally provable in $\mathcal{F}$ *iff* there is a finite sequence of expressions,

$$\psi_1, \psi_2, \psi_3, \ldots, \psi_n,$$

of which it is the terminal element, i. e., $\varphi \equiv \psi_n$, and such that each ψ_i is either an axiom, $\alpha \in \Sigma$, or results from the application of a rule, $\rho \in R$, to earlier expressions in said sequence. We write '$\vdash_{\mathcal{F}} \varphi$' *iff* φ is derivable in $\mathcal{F}$ and '$\mathsf{d} \vdash_{\mathcal{F}} \varphi$' to denote a specific derivation d of φ.

3.2 Encoding

Assume given a Gödel numbering of $\mathcal{F}$, *viz.*, an encoding that assigns each basic character of the language $\mathcal{L}$ and its syntactic entities (e. g., terms, expressions) unique natural numbers. We write '$gn(x)$' for the Gödel number assigned to object x. Based on this encoding, define a stock of primitive recursive relations that are true for natural numbers *iff* a certain syntactic fact is true for the objects they encode. In particular assume given a primitive recursive relation $Proof_F(n, m)$ such that,

$$Proof_F(n, m) \quad \textit{iff} \quad n = gn(\mathsf{d} \vdash_{\mathcal{F}} \varphi) \text{ and } \mathsf{m} = gn(\varphi).$$

Consequently,

$$\exists x Proof_F(x, gn(\varphi)) \quad \textit{iff} \quad \vdash_{\mathcal{F}} \varphi.$$

We suppress the bound variable and write,

$$Pr_F(gn(\varphi)) \quad \textit{iff} \quad \vdash_{\mathcal{F}} \varphi.$$

We call '$Pr_F(x)$' a provability predicate. Note that it is at least Σ_1 due to its existential quantifier.

3.3 Representation

We now consider how things look "inside" the formal system $\mathcal{F}$. For this purpose we assume the language of $\mathcal{F}$ to be or to contain the language of arithmetic, *viz.*, we have available (directly or via interpretation) the symbols '$\mathsf{0}$' (zero), 'S' (successor), '$+$' (plus), and '$\times$' (times). The canonical way to represent a natural number 'n' in the formal language $\mathcal{L}$ is via the numeral '$\overline{\mathsf{n}}$:'

$$\overline{\mathsf{n}} :\equiv \underbrace{\mathsf{S} \ldots \mathsf{S}}_{n \text{ times}} \mathsf{0}.$$

By extension, we write for arithmetical terms: $\overline{\mathsf{t}} :\equiv \overbrace{\mathsf{S} \ldots \mathsf{S}}^{t \text{ times}} \mathsf{0}$. Finally, it helps perspicuity to have a special notation for numerals that represent Gödel numbers:

$$\ulcorner \varphi \urcorner :\equiv \underbrace{\mathsf{S} \;\ldots\ldots\; \mathsf{S}}_{gn(\varphi) \text{ times}} \mathsf{0}.$$

Aside. Here and elsewhere we use '$=$' to denote ordinary equality and '$\equiv$' to denote the syntactic identity of strings of symbols; we prefix either one with a colon in case they hold by definition. Single apostrophes have no meaning, they are used as spacing devices to help legibility.

Assume the formal system $\mathcal{F}$ to contain (directly or via interpretation) the axioms of what is called Robinson's Arithmetic $\mathcal{Q}$—*viz.*, the six Peano-Dedekind axioms that define successor, addition, and multiplication plus a seventh axiom stating that any number is zero or a successor. For all such systems we can establish that any recursive fact that is true (or false) can also be derived in $\mathcal{F}$ (or its negation can).

Fact 1. Assume $\mathcal{F}$ to contain $\mathcal{Q}$ and to be consistent. For every recursive relation $R(x_1, \ldots, x_k)$ there is a corresponding formal expression $\varphi_{\mathsf{R}}(\mathsf{x}_1, \ldots, \mathsf{x}_\mathsf{k})$ in the language of $\mathcal{F}$ such that, for all $n \in \mathbb{N}$,

$$\begin{aligned} R(n_1, \ldots, n_k) &\quad\Leftrightarrow\quad \vdash_{\mathcal{F}} \varphi_{\mathsf{R}}(\overline{\mathsf{n}}_1, \ldots, \overline{\mathsf{n}}_\mathsf{k}), \\ \text{not-}R(n_1, \ldots, n_k) &\quad\Leftrightarrow\quad \vdash_{\mathcal{F}} \neg\varphi_{\mathsf{R}}(\overline{\mathsf{n}}_1, \ldots, \overline{\mathsf{n}}_\mathsf{k}). \end{aligned}$$

Remark. The presence of $\mathcal{Q}$ guarantees the left-to-right direction '$\Rightarrow$' in the two equivalences, consistency their reverse. Note how we use sans-serif font to communicate formal expressions in the language of $\mathcal{F}$.

From Fact 1 follows immediately,

Corollary 1. Assume $\mathcal{F}$ to be as in Fact 1. For the primitive recursive proof relation $Proof_F(x, y)$ there is a corresponding formal expression $\mathsf{Proof}_\mathsf{F}(\mathsf{x}, \mathsf{y})$ in the language of $\mathcal{F}$ such that, for all $n \in \mathbb{N}$,

$$\begin{aligned} Proof_F(n, m) &\quad\Leftrightarrow\quad \vdash_{\mathcal{F}} \mathsf{Proof}_\mathsf{F}(\overline{\mathsf{n}}, \overline{\mathsf{m}}), \\ \text{not-}Proof_F(n, m) &\quad\Leftrightarrow\quad \vdash_{\mathcal{F}} \neg\mathsf{Proof}_\mathsf{F}(\overline{\mathsf{n}}, \overline{\mathsf{m}}). \end{aligned}$$

By adding the existential quantifier to the mix, we can extend these results and arrive at Σ_1-completeness and Σ_1-soundness of the formal system $\mathcal{F}$ (i. e., completeness and soundness in respect to all expressions, recursive or formal, with at most one existential quantifier in front).[1] This then allows

[1] The situation is more complex than it was before since simple consistency does no longer suffice as it did for Fact 1 to secure the direction from right-to-left, i. e., soundness; or, in the current case, Σ_1-soundness. This is the reason why Gödel introduced the concept of ω-consistency, which Rosser circumvented again by building it right into the provability predicate. We ignore these issues here; the curious reader may turn to (Buldt 2014, § 2.1).

us to reason as follows:

$$\begin{array}{cl@{\qquad\qquad}cl} \vdash_{\mathcal{F}} \varphi & & \vdash_{\mathcal{F}} \varphi & \\ \Downarrow & \text{; by definition of } Pr(x) & \Uparrow & \text{; by definition of } Pr(x) \\ Pr_F(gn(\varphi)) & & Pr_F(gn(\varphi)) & \\ \Downarrow & \text{; by } \Sigma_1\text{-completeness} & \Uparrow & \text{; by } \Sigma_1\text{-soundness} \\ \vdash_{\mathcal{F}} \mathsf{Pr}_{\mathsf{F}}(\ulcorner\varphi\urcorner) & & \vdash_{\mathcal{F}} \mathsf{Pr}_{\mathsf{F}}(\ulcorner\varphi\urcorner) & \end{array}$$

which gives us as a corollary the *a*dequacy (i. e., completeness and soundness) of the *f*ormal system in respect to a *f*ormalized *p*rovability *p*redicate:

Corollary 2. (AFP) $\vdash_{\mathcal{F}} \varphi \ \Leftrightarrow \vdash_{\mathcal{F}} \mathsf{Pr}_{\mathsf{F}}(\ulcorner\varphi\urcorner)$.

4 Gödel's Proof

Assume a formal system to satisfy,

$$\text{(AFP)} \quad \vdash_{\mathcal{F}} \varphi \ \Leftrightarrow \vdash_{\mathcal{F}} \mathsf{Pr}_{\mathsf{F}}(\ulcorner\varphi\urcorner),$$

and the fixed point equivalence (FPE), *viz.*, the existence of a sentence $\gamma \in \mathcal{L}$ such that,

$$\text{(FPE)} \quad \vdash_{\mathcal{F}} \gamma \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner).$$

The formal undecidability of the fixed point γ, and hence the incompleteness of the formal system $\mathcal{F}$, follows then immediately from the properties of γ as per FPE in conjunction with AFP. The proof (sketched below) assumes $\mathcal{F}$ to be consistent (indicated by "$\mathsf{con}\ \mathcal{F}$" where it kicks in) and is by contradiction (indicated by the symbol "$\lightning$" when reached).

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

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

For our purposes this sketch is sufficiently close to Gödel's original proof. The fixed point equivalence, FPE, is what the remainder of the paper is about. Before we can proceed, however, a clarification seems in order.

There are many different ways to establish Gödel's incompleteness result. Some are fairly robust (i. e., transfer easily from one system to another), some are tailored for one specific formal system. For example, there are model theoretic proofs that require the absence of induction while others require its presence. This raises the question of what proof exactly we refer to when we speak of "Gödel's incompleteness proof." The two extreme options "all proofs" and "Gödel's original proof" are not helpful for obvious reasons.

There are two features that distinguish Gödel's original proof. First, specifications for the formal system $\mathcal{F}$ are known to be demonstrably minimal: $\mathcal{F}$ must allow for representation and be somewhat sound (i. e., what we summed up in AFP). Making minimal demands results in a proof that applies to the widest possible class of formal systems where it matters as little as possible whether the language is rich or poor, the rules are strong or weak, the axioms powerful or not. We shall say, "the proof scales well," to catch this property. Second, the proof itself relies on fixed points (i. e., FPE above). In Gödel's original proof this made the formally undecidable sentence to have lowest quantificational complexity possible, *viz.*, it was Π_1 (i. e., one universal quantifier in prenex normal form). We shall say, "the proof is optimal," when this is the case.

Guided by those two properties, we call every proof P, which scales well and is optimal, "Gödel's incompleteness proof." The recoined phrase now designates an informally characterized equivalence class, if you will. And the justification for calling it so is that all proofs in this class actually proceed by minor variations of the proof sketched above (see Buldt 2014, § 2). This is why the following observations apply to the entire class.

5 Proof of the Fixed Point Equivalence

Gödel's original fixed point construction proceeded in three step (roughly) as follows.

Step 1: Substitution

- Fix a certain individual variable of your choice; say 'u.'
- Define a substitution function *sub* that mirrors the substitution of the replacee variable 'u' for a replacer term 't,'

$$\varphi[\mathsf{u}]\tfrac{\mathsf{t}}{\mathsf{u}} \equiv \varphi(\mathsf{t}),$$

but in the realm of Gödel numbers. In short:

$$sub(x,y) := \begin{cases} gn(\varphi[\mathsf{u}]\tfrac{\bar{\mathsf{t}}}{\mathsf{u}}) & \text{if } x = gn(\varphi(\mathsf{u})) \text{ and } y = gn(\bar{\mathsf{t}}) \\ x & \text{otherwise.} \end{cases}$$

- Note that $sub(x,y)$ is primitive recursive and therefore represented in $\mathcal{F}$ by an expression $\varphi_{\mathsf{s}}(\mathsf{x},\mathsf{y})$.

Step 2: Definitions

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

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

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

Step 3: Derivation.

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

$$\begin{array}{lll} \vdash_{\mathcal{F}} \neg\mathsf{Pr}_{\mathsf{F}}(\mathsf{x}) \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\mathsf{x}) & ; & \text{logic} \\ \vdash_{\mathcal{F}} \neg\mathsf{Pr}_{\mathsf{F}}(\mathsf{sub}(\overline{\mathsf{p}},\overline{\mathsf{p}})) \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner) & ; & \text{Step 2} \\ \vdash_{\mathcal{F}} \forall\mathsf{x}\big[\neg\mathsf{Proof}_{\mathsf{F}}(\mathsf{x},\mathsf{sub}(\overline{\mathsf{p}},\overline{\mathsf{p}}))\big] \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner) & ; & \text{def. } \mathsf{Pr}_{\mathsf{F}} \\ \vdash_{\mathcal{F}} \varphi(\overline{\mathsf{p}}) \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner) & ; & \text{def. } \varphi(\overline{\mathsf{p}}) \\ \vdash_{\mathcal{F}} \gamma \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner) & ; & \text{def. } \gamma \\ & & \square \end{array}$$

Remark. Gödel assumed a formal system in which induction is available. This can be used to prove that expressions $\varphi_{\mathsf{f}}(\vec{\mathsf{x}}) \in \mathcal{L}$, which represent a function $f(\vec{x})$ in $\mathcal{F}$, can actually be brought into the functional normal form of terms $\varphi_{\mathsf{f}}(\vec{\mathsf{x}}) = \mathsf{y}$ which is presupposed in step 2 of the derivation. One can do without this convenience, but the argument, although it remains the same, gets cluttered by longer expressions and longer derivations. The same is true for Tarski's suggestion to use a diagonalization function $diag(x) := gn(x(\ulcorner\mathsf{x}\urcorner))$, with $x = gn(\varphi(\mathsf{u}))$—which amounts to skipping the substitution function because $sub(x,x) = diag(x)$.

The proof seems to work by black magic (a common joke about its twin, the recursion theorem), and sometimes even seasoned logicians admit to being baffled by and therefore not liking it. We wish to show that this is a misapprehension and that under closer scrutiny the magic turns out to be nothing but mundane diagonalization. But before we can do that, we need to rehash diagonalization.

6 Diagonalization

Diagonalization may mean slightly different things in different contexts. The primary meaning for our purpose derives from arranging an at most countable set of objects, $A = \{a_0, a_1, a_2, \ldots\}$, in a two-dimensional array, $\mathcal{A} = \{a_{ij}\}_{i,j \in \omega}$, according to rows, R_i, and columns, C_j, as follows:

$$\begin{array}{cccccc} & C_0 & C_1 & & C_n & \\ R_0 & a_{00} & a_{01} & \dots & a_{0n} & \dots \\ R_1 & a_{10} & a_{11} & \dots & a_{1n} & \dots \\ & \vdots & \vdots & \ddots & \vdots & \\ R_n & a_{n0} & a_{n1} & \dots & a_{nn} & \dots \\ & \vdots & \vdots & & \vdots & \ddots \end{array}$$

Note that the double index "ij," which indicates the location on the grid, is different from the index used previously in "$A = \{a_0, a_1, a_2, \ldots\}$" to differentiate otherwise unspecified objects from one another. In other words, in the array above we suppress the 'k' as in '$a_{k_{ij}}$.'

We call a function $f : A \to A$ a sequence transforming function and write 'F' when it is applied to each element of a sequence $S = \langle a_0, \ldots, a_n, \ldots \rangle$ and thus transforms S into the sequence $F(S) = \langle f(a_0), \ldots, f(a_n), \ldots \rangle$. Furthermore, we call the sequence $\langle a_{00}, a_{11}, a_{22} \ldots, a_{nn}, \ldots \rangle$ in an array $\mathcal{A}$ the diagonal and abbreviate it 'D.'

Let an array $\mathcal{A}$ be given and F be a sequence transforming function. Apply F to the diagonal D in $\mathcal{A}$ to get what is called the anti-diagonal D'.

$$D' := F(D) = \langle f(a_{00}), f(a_{11}), f(a_{22}), \ldots, f(a_{nn}), \ldots \rangle.$$

One out of two things can happen.

1. D' is identical to one of the rows in $\mathcal{A}$; *viz.*, $F(D) = R_i$, for some i such that $R_i \in \mathcal{A}$.

2. D' is not identical to any of the rows in $\mathcal{A}$; *viz.*, $F(D) \neq R_i$, for all i such that $R_i \in \mathcal{A}$.

If the latter case applies, we say that $\mathcal{A}$ is not closed under F, and what we have is Cantor's diagonal argument showing that a certain sequence is not in $\mathcal{A}$. In other words, we have "diagonalized out" (*sc.*, of $\mathcal{A}$). We will not be interested in this case.

If, however, what obtains is the first case and D' is identical to one of the rows, *viz.*, for some i: $F(D) = R_i$, then $\mathcal{A}$ closed under F and F has fixed points. To see this, first observe that the identities $D' = F(D) = R_i$ are defined as element-wise identity. That is, $D' = R_i$ *iff*

$$f(a_{00}) = a_{i0},\ f(a_{11}) = a_{i1},\ \ldots,\ f(a_{ii}) = a_{ii},\ \ldots,\ f(a_{nn}) = a_{in},\ \ldots$$

or, depicted more visually,

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

If we highlight, in $\mathcal{A}$, the anti-diagonal D' as the row R_i it is identical with, then it becomes apparent that F must have (at least) the fixed point $f(a_{ii}) = a_{ii}$.

$$\begin{array}{r|ccccccc} & C_0 & C_1 & & C_i & & C_n & \\ R_0 & a_{00} & a_{01} & \ldots & a_{0i} & \ldots & a_{0n} & \ldots \\ R_1 & a_{10} & a_{11} & \ldots & a_{1i} & \ldots & a_{1n} & \ldots \\ & \vdots & \vdots & \ddots & \vdots & & \vdots & \\ F(D) = R_i & \frac{fa_{00}}{=a_{i0}} & \frac{fa_{11}}{=a_{i1}} & \ldots & \frac{fa_{ii}}{=a_{ii}} & \ldots & \frac{fa_{nn}}{=a_{in}} & \ldots \\ & \vdots & \vdots & & \vdots & \ddots & \vdots & \\ R_n & a_{n0} & a_{n1} & \ldots & a_{ni} & \ldots & a_{nn} & \ldots \end{array}$$

We now have the prerequisites to show that $\gamma \leftrightarrow \neg\mathsf{Pr}_{\mathsf{F}}(\ulcorner\gamma\urcorner)$ is an instance of $f(a_{ii}) = a_{ii}$ for suitable chosen $\mathcal{A}$ and F.

7 Fixed Points and Diagonalization

According to the previous section, all we need to do is find suitable choices for $\mathcal{A}$ and F.

We construct the set of objects, which to lay out in the array $\mathcal{A}$, in three steps. In Step 1 we choose all first-order expressions with the free variable 'u:'

$$A = \{\varphi_0(\mathsf{u}), \varphi_1(\mathsf{u}), \varphi_2(\mathsf{u}), \ldots\}.$$

In Step 2 we form the set of all of their Gödel numbers:

$$B = \{\ulcorner\varphi_0(\mathsf{u})\urcorner, \ulcorner\varphi_1(\mathsf{u})\urcorner, \ulcorner\varphi_2(\mathsf{u})\urcorner, \ldots\}.$$

In Step 3 we systematically plug all members of B into the free variable slots of all members of A. Let $\ulcorner\varphi_b\urcorner$ denote an arbitrary element of B and $\varphi_a(\mathsf{u})$ an arbitrary element of A; then our new set C consists of all elements $\varphi_a(\ulcorner\varphi_b\urcorner)$, for all a, b. We write 'φ_{ab}' instead of '$\varphi_a(\ulcorner\varphi_b\urcorner)$.'

The set C is what we construct our array $\mathcal{A}$ from. We 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 in the fixed point construction of Section 5. We may call this the "first diagonalization."

Next up is to find a suitable transformation F.

(1) Observe that the provability predicate $\neg\mathsf{Pr}_{\mathsf{F}}(\mathsf{u})$ is itself a member of the first set we started out with: $A = \{\varphi_0, \varphi_1, \varphi_2, \ldots\}$. Thus, there is an index i such that $\varphi_i \equiv \neg\mathsf{Pr}_{\mathsf{F}}(\mathsf{u})$.

(2) Apply the transformation $f : \varphi_{ab} \mapsto \neg\mathsf{Pr}_{\mathsf{F}}(\varphi_{ab})$. Because of (1), f maps C onto C, C will be closed under f (and $\mathcal{A}$ under F), and each image $\neg\mathsf{Pr}_{\mathsf{F}}(\varphi_{ab})$ must be a φ_{in}, for some n. (The first index i is the one from (1).)

(3) We saw in the previous section that closure implies fixed points. Hence, $F(D)$ has a fixed point φ_{ii}. Inspection then shows that φ_{ii} corresponds to the expression $\gamma \equiv \varphi(\overline{\mathsf{p}})$ we used in the fixed point construction of Section 5. We may call this the "second diagonalization."

We thus find that fixed points in formal systems of arithmetic are not the result of black magic nor do they defy our intuitive grasp. Rather, fixed points owe their existence to the fact that certain sets, such as C, are closed under suitable transformations. And the entire procedure—basically, a double (first and second) diagonalization—is entirely syntactic in nature and at no point employs a semantic concept.

8 Varieties of Self-Reference

The original sin was committed by Gödel (1931) himself when he, trying to ease readers into accepting the paper's results and provide them with an intuitive grasp, wrote, "We therefore have before us a proposition that says about itself that it is not provable." But it is difficult to see how a Gödel sentence γ, which satisfies the fixed point equivalence,

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

is supposed to accomplish that. For all FPE states is a truth-functional equivalence. But any two true arithmetical sentences, say, '$1+1 = 2$' and '$2\times 3 = 6$,' are truth-functionally equivalent without implying anything about what they "say," *viz.*, their meaning. In short, truth-functional equivalence is a far car from meaning equivalence. We should also recall the fact that a fixed point such as γ is not a uniquely determined expression but just one among infinitely many equivalent fixed points. This is the price we pay for working with an extensional logic that cares about one thing and one thing only, truth values. Despite these reservations, we briefly discuss two approaches that were thought to have some promise for speaking about self-reference.

8.1 Propertual self-reference

We use the mock word "propertual"—meaning, "pertaining to properties"—to have a catchy counterpart to the more established adjective "objectual," which the next subsection will be about. (I considered calling it semantic and syntactic self-reference but, while not completely off, felt those two not quite capture what I wanted them to convey.) Thus, propertual self-reference occurs when a sentence expresses a property it itself has. It seems straightforward to work this idea out.

Step 1. Identify a property $\mathcal{P}$ with its corresponding set $P = \{x : \mathcal{P}(x)\}$, and identify $\mathcal{P}$'s definition (*sc.*, relative to a structure $\mathfrak{A}$) with a suitable open formula $\varphi(\mathsf{u})$ that is true (*sc.*, in $\mathfrak{A}$) exactly for the elements of P. Let '$\#\mathsf{x}$' denote a name of x in the language of $\mathfrak{A}$. Then we have,

$$\varphi(\mathsf{u}) \ \mathfrak{A}\text{-defines } \mathcal{P} \quad \textit{iff} \quad x \in P \ \Leftrightarrow \ \mathfrak{A} \models \varphi(\#\mathsf{x}).$$

Step 2. Adopt a naming convention, which somehow derives a property's name from its definition, and suppose names to be terms in the language. Regarding the above sketch, for instance, the Gödel number of the defining expression would fit the bill. Let '$\#\varphi(\mathsf{u})$' be such a name. Then we can define propertual self-reference as follows.

Definition. (Propertual Self-Reference) *It is an instance of propertual self-reference iff*

1. $\varphi(\mathsf{u})$ *is an expression, which* $\mathfrak{A}$*-defines property* $\mathcal{P}$*, and*
2. $\mathfrak{A} \models \varphi(\#\varphi(\mathsf{u}))$.

In such cases we say the expression '$\varphi(\#\varphi(\mathsf{u}))$' is (propertually) self-referential, or, to have the property it expresses.

Step 3. Adopt a naming convention that allows Gödel numbers to be names. Let $\mathfrak{N}$ be the standard model of arithmetic and observe that $\neg\mathsf{Pr}_\mathsf{F}(\mathsf{u})$ $\mathfrak{N}$-defines unprovability in $\mathcal{F}$. Then it is easy to see that, since $\nvdash_\mathcal{F} \neg\mathsf{Pr}_\mathsf{F}(\mathsf{u})$, the expression '$\neg\mathsf{Pr}_\mathsf{F}(\ulcorner\neg\mathsf{Pr}_\mathsf{F}(\mathsf{u})\urcorner)$' is self-referential in the sense just defined.

The problem with this idea in our context is that no one knows how to tie all this to fixed points. Suppose we have a fixed point equivalence,

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

whose right-hand side, φ_ψ, can be shown to be propertually self-referential like we did for $\neg\mathsf{Pr}_\mathsf{F}(\ulcorner\neg\mathsf{Pr}_\mathsf{F}(\mathsf{u})\urcorner)$. What conditions would elevate ψ from being merely truth-functionally equivalent (*cf.* the cautionary remarks made at the beginning) 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).

8.2 Objectual self-reference

We arrive at a weaker notion of self-reference when we consider the requirement to express properties as excessive and superfluous baggage and instead just ask whether or not it is permissible for an expression to contain its own name: $\varphi(\#\varphi)$. Permissible could here be understood in various ways, say, as to whether $\varphi(\#\varphi)$ is well-formed, or derivable, or true (which may already cross a line—note our intention to force objectual self-reference to be an almost syntactic notion). In short, can it happen that an expression references itself as an object by containing its own name?

Definition. (Direct Objectual Self-Reference) *It is an instance of direct objectual self-reference iff an expression* φ *contains its own name:* $\varphi(\#\varphi)$.

Instances that exemplify this kind of self-reference are Quine's norm, where an expression φ gets concatenated with its own quotation $|\varphi|$ (here considered to be its name), *viz.*, $\varphi^\frown|\varphi|$, or expressions that contain their own Gödel number: $\varphi(\ulcorner\varphi\urcorner)$.

Like before, it is straightforward to check whether our new concept applies to the fixed point γ from FPE. Recall that γ is shorthand for $\forall \mathsf{x}[\neg\mathsf{Proof_F}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))]$, with $p = gn(\neg\mathsf{Pr_F}(\mathsf{sub}(\mathsf{u}, \mathsf{u}))$. Thus, no direct objectual self-reference occurs in γ.

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. This, relaxing standards, is what we try next. Let '$\#\#\varphi$' denote an indirect name of φ. We say a little more on indirect names below; for the time being think of them as definite descriptions or functional expressions that, once they were evaluated, serve as a name.

Definition. (Indirect Objectual Self-Reference) *It is an instance of indirect objectual self-reference iff an expression φ contains its own indirect name: $\varphi(\#\#\varphi)$.*

Since $sub(\overline{\mathsf{p}}, \overline{\mathsf{p}}) = gn(\gamma)$, the expression γ, which is $\forall \mathsf{x}[\neg\mathsf{Proof_F}(\mathsf{x}, \mathsf{sub}(\overline{\mathsf{p}}, \overline{\mathsf{p}}))]$, contains an indirect name of itself. Thus, γ is self-referential in the indirect objectual sense.

Digression. Some (e. g., Heck 2007) are perfectly happy to embrace the last point and call the Gödel sentence γ self-referential and have it say "I'm not provable." I am not prepared to do that, and here is why: (1) γ does not say "I" but refers to itself indirectly via a functional expression; (2) γ is true *iff* γ is not formally provable. By itself, (2) is a raw datum about γ's model theoretic evaluation and the resulting truth value. As such, it is just another equivalence (see what we said above, pp. 12f.) that implies nothing about meaning or self-reference. And it doesn't imply anything mostly because we lack a theory of self-reference that would allow us to go a step further and interpret this as an instance of self-reference. We have a hunch that it is, but gut feelings are a poor substitute for an actual theory. Likewise for (1).

Let me further illustrate this point. My cat naps. By itself, this is a raw datum. I can add, "the cat's tired, that's why." But this is a cooked up story unless I can back it up with actual science on pet behavior that goes beyond my anthropocentric habit of ascribing pets intentional states. Similar to our intentional stance towards pets (Dennett), all we currently have on self-reference are deeply engrained semantic practices. But in both cases we lack actual theories. The main challenge as I see it is therefore as follows. All you can see from within the formal system are truth values, but no meaning and no self-reference. It is like inside the Matrix, just 0's and 1's, or inside Searle's Chinese Room, just rules-based symbol pushing. But as humans we live and breathe meaning (whatever that is) and by our own design (arithmetization) we can assign numbers a meaning they may have for us; for instance, that they encode something. But like ascribing pets intentional states does not mean they have them, ascribing numbers self-reference does not mean they have it. In other words, teaching a formal system self-reference is almost like teaching the Chinese Room to think. Ignoring this lacuna runs the risk of taking things to be understood that are not and of not developing a theory where one is needed. (*End of digression*)

The naming business presented above was admittedly severely deficient and in need of a much more rigorous treatment. It was based on an oversimplifying account of how naming works in (some?) natural languages: proper names refer directly while definite descriptions need some processing first. Quite obviously, my goal was not to develop a theory. I just wanted to put some quick stakes (i. e., examples) in the ground to show the lay of the land.

To me, the interesting thing about indirect objectual self-reference is not whether it results from fixed point construction (it depends) or whether it offers a promising general approach to self-reference (I doubt it). On the contrary, the interesting thing to me is the exact opposite, namely, that it can be used to construct fixed points.

8.3 Fixed points via self-reference

This approach, first suggested by Quine and subsequently elaborated by Smullyan, uses the norm function. In what follows, all objects are strings of symbols. We write $a^\frown b$ for the concatenation of two strings a, b. Strings fall into two distinct classes; expressions 'a' and their quotations '$|a|$.' Call the *norm of* a the concatenation of a with its own quotation: $a^\frown|a|$.

The intuition behind these definitions is what their names suggest. Thus, if, "$a \equiv$ don't quote me on," then, "$|a| \equiv$ 'don't quote me on'," and a's norm is, "$a^\frown|a| \equiv$ don't quote me on 'don't quote me on'."

Thus, objectual self-reference is not a rare occurence but built right into the mechanics due to the presence of quotation. Ordinary self-reference is easily achieved, too. Let a be "this is the norm of." Then its norm is: "this is the norm of 'this is the norm of'." But the first "this" refers to $a^\frown|a|$, which is indeed the norm of "this is the norm of," and the norm refers to itself. Other monstrosities come easy as well; truth-tellers, for example. Let E (a set) and d (an expression) be as follows:

$$d \equiv E \text{ contains the norm of } {}^\frown|E \text{ contains the norm of}|.$$

Observe that d states that the norm of "E contains the norm of" is in E; but that norm is d itself and hence $d \in E$ if d is true. On the other hand, if d is false, then $d \notin E$. Thus, d is true *iff* $d \in E$, and d is a truth-teller for E (i. e., true *iff* in).

Note how the last example used the language of sets. This is crucial for a smooth implementation and requires, in case of first-order theories, to extend their language with class symbols. Doing without an extended language is possible but requires a modified concept of normalization (what Smullyan calls "near normalization"), which, then, is sensitive to the underlying set of basic symbols (e. g., it won't work with the Sheffer stroke). All these complications are the reason not to go into any more detail.

Be the details as they may, the important point was to mention that objectual self-reference as it is embodied in the norm operation can be used to construct fixed points that satisfy FPE.

9 Concluding Remarks

We set out to clarify the respective roles fixed points, diagonalization, and self-reference play in Gödel's proof (properly characterized, see § 4). Accordingly, we highlighted the crucial role fixed points play (§ 4) and then disenchanted the alleged mystery of their existence by showing that they result—and quite naturally so?—from a double diagonalization (§ 5, § 7). Our treatment made clear, or so I hope, that nothing semantic enters the picture and that self-reference in particular is totally absent. We then turned to the concept of self-reference and concluded that as of now we do not have at our disposal a theory that would allow us to reconstruct and justify the intuitive hunches we have on this topic (§ 8.1–2). To round things off, we mentioned in § 8.3 that in the disguise of normalization a syntactic form of objectual self-reference can be used to construct fixed points. Taking it all together, the highlighted interdependencies can be visualized as follows:

$$\begin{tikzcd}
\text{diagonalization} \arrow[r, Rightarrow] & \text{fixed points} \arrow[d, Rightarrow] & \text{(obj.) self-reference} \arrow[l, Rightarrow] \\
& \text{incompleteness} &
\end{tikzcd}$$

What, then, one might be tempted to ask, is the crucial player in this picture? We need to back up a bit in order for me to say where I believe a good answer can be found.

Let $\mathbf{2} = \{0, 1\}$ and $\mathbf{2}^{\mathbb{N}}$ the set of all functions $f : \mathbb{N} \to \mathbf{2}$. Recall that $\mathbf{2}^{\mathbb{N}}$ has the cardinality of the powerset of $\mathbb{N}$. Cantor used his famous diagonal argument to prove that there is no surjective function $f : \mathbb{N} \to \mathbf{2}^{\mathbb{N}}$. Lawvere gave a category theoretic generalizations of this argument that, boiled down again to the level of set theory, establishes roughly the following claim. Let T and Y be sets and Y^T be the set of functions from T to Y. Then, if Y is appropriately well-behaved, there is no onto function $g : T \to Y^T$ ("diagonalizing out"). Or, in its contraposed form, if there is such an onto function g, then all functions $h : Y \to Y$ have a fixed point. Yanovsky (2003) shows how all the usual suspects (i. e., paradoxes and limitative theorems) can be couched in terms of this framework and then follow from the generalized Cantor theorem.

To help us wrap our brains around the result Yanovsky offers the following crutch. Note that $\mathbf{2}^{\mathbb{N}}$ is the set of all characteristic functions and thereby a set of properties elements of $\mathbb{N}$ may have. Similar for Y^T and T. The two versions of the theorem leave us hence with a choice: either the set of properties is well-behaved but suffers limited expressibility, or properties are completely expressible but show degenerated behavior in terms of fixed points—an undesirable outcome either way.

This more general picture, if accurate, seems to align well with our much more limited analysis: both put diagonalization in the driver's seat—where Wolfgang Spohn has been for the better part of his career.

References

Buldt, Bernd. 2014 ≫The scope of Gödel's first theorem.≪ *Logica Universalis* 8 (3–4):499–552.

Halbach, Volker and Visser, Albert. 2014 ≫Self-reference in arithmetic I+II.≪ *Review of Symbolic Logic* 7 (4):671–691, 692–712.

Heck, Richard. 2007 ≫Self-reference and the languages of arithmetic.≪ *Philosophia Mathematica* (III) 15 (1):1–29.

Yanofsky, Noson Y. 2003 ≫A universal approach to self-referential paradoxes, incompleteness and fixed points.≪ *Bulletin of Symbolic Logic* 9 (3):362–386.