Incompleteness via paradox and completeness

Walter Dean
Department of Philosophy
University of Warwick
e-mail: W.H.Dean@wawick.ac.uk

[*Gödel's incompleteness theorem*] *is by no means to be judged as only a negative result, rather it plays a similar role for proof theory as does the discovery of the irrational numbers for arithmetic.* Bernays (1954b), p. 11

Wittgenstein simply did not know what to say about the paradoxes. I don't either. But one thing is clear: *the fruitful problem is not to 'get rid of them' but to get something out of them.* Kreisel (1958), p. 157

Abstract. This paper explores the relationship borne by the traditional paradoxes of set theory and semantics to formal incompleteness phenomena. A central tool is the application of the Arithmetized Completeness Theorem to systems of second-order arithmetic and set theory in which various "paradoxical notions" for first-order languages can be formalized. I will first discuss the setting in which this result was originally presented by Hilbert & Bernays (1939) and also how it was later adapted by Kreisel (1950) and Wang (1955) in order to obtain formal undecidability results. A generalization of this method will then be presented whereby Russell's paradox, a variant of Mirimanoff's paradox, the Liar, and the Grelling-Nelson paradox may be uniformly transformed into incompleteness theorems. Some additional observations are then framed relating these results to the unification of the set theoretic and semantic paradoxes, the intensionality of arithmetization (in the sense of Feferman, 1960), and axiomatic theories of truth.

§1. Introduction The goal of this paper is to highlight by both mathematical and historical means the relationship borne by the traditional paradoxes of set theory and semantics to formal incompleteness phenomena. This is by no means a novel theme. Perhaps most famously, Gödel (1931b) remarked that the analogy between the proof of his first incompleteness theorem and the Richard paradox "leaps to the eye". He then commented in a footnote that "any epistemological antinomy could be used for a similar proof of the existence of undecidable propositions" (1931b, p. 149) as well as elaborating on the analogy between his first incompleteness theorem and the Liar paradox in his correspondence with Zermelo (Gödel, 1931a).[1]

Gödel's prediction has been borne out by the work of many subsequent authors who have obtained formal undecidability results by formalizing a variety of other paradoxes which we would now classify as "semantic" but which would have been termed "epistemic" according

2010 *Mathematics Subject Classification*: 03C62, 03H15, 03C57, 00A30.
Keywords and phrases: Arithmetized Completeness Theorem, interpretability, set theoretic paradoxes, semantic paradoxes, Hilbert program, David Hilbert, Paul Bernays, Georg Kreisel, Hao Wang.

[1] In addition to the introductory remarks to (Gödel, 1931a,b), see also (Wang, 1981, pp. 6-8) for further discussion of Gödel's engagement with the paradoxes *en route* to the incompleteness theorems.

to the original terminology of Ramsey (1926) employed by Gödel. An early waypoint here was Robinson's (1963) use of Berry's paradox to prove Tarski's theorem on the definability of truth in a manner which anticipates its later use by Vopěnka (1966), Boolos (1989), and Kikuchi (1997) to prove incompleteness theorems. More recent applications of a similar sort include formalizations of the Grelling-Nelson paradox by Cieśliński (2002) and Kripke (2014), the surprise exam paradox by Kritchman & Raz (2010), and Yablo's paradox by Priest (1997b).[2]

What is less well known is how these results are related to the set theoretic paradoxes and also how they relate to the line of research which culminated in the two volumes of Hilbert & Bernays's *Grundlagen der Mathematik* (1934; 1939). This work is notable in the present context for two reasons. First, it contains an extended discussion of the paradoxes, both as an introduction to and in light of Gödel's incompleteness theorems. Second, the *Grundlagen* also contains a formalization of Gödel's (1930) completeness theorem for first-order logic in the form of what has come to be known as the *Arithmetized Completeness Theorem*. A simple form of this result states that if a purely relational formula of first-order logic cannot be refuted from the axioms of the predicate calculus, then it is satisfiable in an *arithmetical model* — i.e. one which not only has domain $\mathbb{N}$ but is also such that all of its non-logical symbols are interpreted by first-order arithmetical formulas.

As I will discuss further in §2, Hilbert & Bernays first presented the paradoxes by treating semantic notions such as truth and denotation as primitives and then showing how the adoption of natural principles leads to inconsistency relative to an appropriate background theory. This approach is notable in part due to its similarity to contemporary axiomatic approaches to truth. But Hilbert & Bernays also suggested that the reasoning of the Liar provides a template for the proof of the first incompleteness theorem. They then went on to formulate a second-order truth definition for first-order arithmetic, observed that this definition can be used to provide a consistency proof for a system similar to PA, before finally remarking that such a proof cannot be regarded as finitary.

These observations were combined in a novel but largely overlooked way by Georg Kreisel (1950; 1953) and Hao Wang (1955) who observed it is also possible to formulate a truth definition for first-order arithmetic in a system S similar to what is now called *Gödel-Bernays set theory*. By applying the Arithmetized Completeness Theorem to S, Kreisel and Wang obtained a first-order arithmetical interpretation of this theory relative to its formal consistency statement. They then observed that the images under such an interpretation of paradoxical statements resembling the Liar or the assertion that the Russell class is not a member of itself are formally undecidable in S.[3]

After introducing some preliminaries in §3, the goal of §4 below will be to provide a modern reconstruction and generalization of this method for obtaining formal incompleteness results. Part of the interest of such a project derives from the manner in which Kreisel and Wang's work is juxtaposed between the foundational concerns of the 1920s and 1930s and the subsequent development of mathematical logic from the 1950s onward.[4] More generally, I will also suggest that the undecidable statements obtained in the manner

---

[2] See (Kotlarski, 2004) and (Kikuchi & Kurahashi, 2016) for further surveys of results in this vein.

[3] Kreisel (1968, p. 382, note 43) later sketched a proof of Gödel's second incompleteness theorem based on a related construction. This was subsequently popularized by Smorynski (1977, §6) and is studied in greater detail by Kikuchi & Tanaka (1994) and Manevitz & Stavi (1980). The second of these treatments illustrate several complexities involved with iterating Kreisel's construction which will not be relevant here until §5.5.

[4] Additional motivation for focusing on the results in their original context derives from the opacity of Kreisel's original presentation in (1950). For as Wang subsequently observed, this paper not only contains a number of elisions and at least one apparent error, it also is "excessively condensed so that the reader would have to reconstruct

in question are of independent interest. For not only are their formal properties distinct from those obtained in the manner of Gödel's (1931b) original proof or in the manner of subsequent combinatorial independence results — e.g. in virtue of being Δ^0_2 rather than Π^0_1 or Π^0_2 — but they are also generated in a different manner. For rather than starting out by replacing putatively paradoxical notions like truth or denotation with surrogates like provability or demonstrable denotation, the method introduced by Kreisel and Wang begins with the observation that the relevant semantic notions for first-order languages are *definable* in appropriate higher-order extensions.

The independent statements obtained in this manner thus bear a closer resemblance to the putatively paradoxical ones than do those which figure in Gödel's original proof or in (e.g.) Boolos's (1989) formalization of Berry's paradox. As a consequence, the hope remains that results obtained in the manner of Kreisel and Wang still have something to teach us about the paradoxes which were used to generate them. Although both authors hinted at such a possibility, the thought that formal incompleteness might play a role in a uniform response to the paradoxes has also been largely overlooked by subsequent authors. I will develop this theme further in §5 in the context of describing how the method in question is related to the unification of the set theoretic and semantic paradoxes, absolute undecidability, the intensionality of arithmetization, and axiomatic theories of truth before offering a brief synthesis in §6.

§2. Historical setting Building on his earlier work in geometry (1899), Hilbert first employed the method of arithmetical interpretations to logical systems in his 1917–1918 lectures on the foundations of mathematics (1917–1918). In this setting, such an interpretation takes the form of an assignment of the numerical values 0 (true) and 1 (false) to the propositional variables $X, Y, Z, \ldots$, together with the interpretation of disjunction as multiplication and negation as $1 - X$. A formula is taken to be *valid* if it evaluates to 0 under all such interpretations. Hilbert used these definitions to show the consistency and completeness of his axioms for the propositional calculus.[5]

In the first edition of Hilbert and Ackermann's textbook the *Grundzüge der Theoretischen Logik* (1928) a related technique is described for transforming first-order formulas into propositional ones, whereby each predicate symbol is associated with a collection of numerical substitution instances derived by replacing universally quantified formulas with conjunctions and existentially quantified ones with disjunctions. This yields a method for constructing counter-models to show the non-validity of formulas falsifiable in a finite domain by showing how they can be replaced by equi-satisfiable propositional formulas and invoking the completeness theorem for propositional logic. Gödel showed in his (1930) completeness proof for first-order logic how this technique could be extended to formulas possessing no finite countermodels by proving a form of what we would now recognize as Herbrand's Theorem and invoking König's Infinity Lemma.[6]

In the second edition of the *Grundzüge* (1938), Hilbert and Ackermann provide an exposition of Gödel's completeness proof which is then further formalized in §4.2 of

---

certain steps for himself" (1953, p. 181). In regard to these points, see notes 18, 30, and 55 below.

[5] More precisely, Hilbert originally showed the *Post completeness* of the propositional axioms —- i.e. if an underivable formula were to be added to the calculus as a scheme, the system would be become inconsistent. This was extended in Bernays's 1918 *Habilitationsschrift* to show semantic completeness in the familiar form "every valid formula is provable". See (Zach, 1999) for further discussion of these developments.

[6] See (Dean, 2017) for further discussion of the historical context of Gödel's proof and its reception by Bernays, the introductory notes to (Feferman et al., 1986) for a detailed reconstruction, and (Dean & Walsh, 2017, §4) for more on the role of König's Lemma.

the second volume of the *Grundlagen* (1939). By this point Hilbert and Bernays had also elaborated substantially on the role of what they referred to as the *method of arithmetization* for the development of the so-called *finitary standpoint* described in the first two chapters of the first volume (1934). As I will discuss further in §5.1, one aspect of their view was that such interpretations may be of use in establishing the consistency of systems of analysis or set theory for which the existence of models otherwise represents an "idealizing assumption that properly augments the assumptions formulated in the axioms" (1934, p. 2/2).

These considerations set the stage for Hilbert and Bernays's arithmetization of Gödel's completeness theorem in the second volume of the *Grundlagen*. Unlike modern presentations, this begins with a careful exposition in §4.1 of Gödel's method of the arithmetization of syntax. This machinery is employed to define a primitive recursive function $q(x)$ for each first-order formula φ such that $q(\overline{n}) = 0$ for all $n \in \mathbb{N}$ if and only if φ is irrefutable from Hilbert's axioms for first-order logic. The proof of the completeness theorem given in §4.2 then proceeds by showing how the arithmetized syntactic definitions can be used to formalize the steps in Hilbert and Ackermann's rendition of Gödel's proof. This leads to a method for transforming an irreftutable formula φ of an arbitrary first-order language into a true arithmetical formula φ^* by replacing its non-logical symbols by arithmetical formulas extracted from Gödel's construction. The proof thus yields one form of what is now called the *Arithmetized Completeness Theorem*:

THEOREM 2.1 *Let φ be a sentence of the first-order predicate calculus containing predicate letters $P_1(\vec{x}), \ldots, P_n(\vec{x})$ of arities $k_1, \ldots, k_n$ such that $q(\overline{n}) = 0$ holds for all $n \in \mathbb{N}$. Then there exist formulas $P_1^*(\vec{x}), \ldots, P_n^*(\vec{x})$ also of arities $k_1, \ldots, k_n$ over the language $\mathcal{L}_{\mathsf{Z}} = \{0, s, +, x, <\}$ such that $\mathcal{N} \models \varphi^*$ where $\mathcal{N}$ denotes the standard model of arithmetic and φ^* denotes the $\mathcal{L}_{\mathsf{Z}}$-formula formed by uniformly substituting P_i^* for P_i in φ.*[7]

The presentation of this result marks the point in the *Grundlagen* at which Hilbert & Bernays transition from their discussion of Gödel's completeness theorem to their discussion of his incompleteness theorems. Their account of the second incompleteness theorem in §5.1c famously provides the first full details of the proof at which Gödel (1931b) had merely hinted. But Hilbert & Bernays's treatment of the first theorem in §5.1b is also notable as it is preceded in §5.1a by an extended discussion of the paradoxes. This account in turn is informed by earlier discussions in Hilbert's (1917–1918) lectures on the foundations of mathematics and the *Grundzüge*. Collectively, these sources provide expositions of Russell's paradox, the Liar, and several paradoxes involving the concept of denotation.

Hilbert is now thought to have learned of Russell's paradox soon after its independent discovery by Zermelo in or around 1899. This marked the beginning of a period during which the mathematical significance of various paradoxes and antinomies was actively discussed in Göttingen. During this period, Hilbert and his collaborators also undertook a careful consideration of Russell's logicism, the later phases of which are reflected in (1917–1918). This presentation served as the basis for §IV.4 of the *Grundzüge* (1928) in which Russell's paradox is presented as a variant of the Grelling–Nelson paradox.[8]

---

[7] This formulation is given on (1939, pp. 252-253/260-263) and has traditionally been attributed to Bernays. A partial reconstruction of the original proof is given by Kleene (1952, §72) which is in turn presented in greater detail by Ebbs (2015). Note that the statement is false if φ is allowed to contain the identity symbol $=$ (for in this case, φ might be satisfiable only in a model of fixed finite size). This restriction may be lifted if identity is treated non-logically such that occurrences of $=$ may be interpreted by an arithmetical formula $=^*$ defining an equivalence relation on the natural numbers.

[8] See (Peckhaus & Kahle, 2002) and (Mancosu, 2003) for discussion of these developments.

Although the *Grundzüge* also contains a discussion of the Liar, this is developed in greater detail in §5.1a of the second volume of the *Grundlagen* relative to an arithmetical background theory which Hilbert & Bernays refer to as *Formalism* F. This system is not presented axiomatically but rather asserted to extend what Hilbert & Bernays refer to as *rekursive Zahlentheorie* by adding axioms defining a least-number operator μ and defining equations for primitive recursive functions sufficient for carrying out the arithmetization of syntax as presented in §4.1.[9] The paradox is presented as a *reductio* of the assumption that there is a predicate $\mathfrak{W}(x)$ definable in the language $\mathcal{L}_{\mathsf{F}}$ which provably satisfies the Tarski biconditionals

$$\mathfrak{W}(\ulcorner\varphi\urcorner) \leftrightarrow \varphi \tag{2.1}$$

for all $\mathcal{L}_{\mathsf{F}}$-sentences φ. Hilbert's lectures and the *Grundzüge* also contain descriptions of a paradox similar to that of Berry (as first described by Russell, 1908). The *Grundlagen* (1939, pp. 262-268/271-277) develops a related paradox, which Hilbert and Bernays liken to that of Richard (1905) based on the assumption that there exists a definable denotation function $\mathfrak{d}(x)$ which provably satisfies

$$\mathfrak{d}(\ulcorner t\urcorner) = \mathfrak{t} \tag{2.2}$$

for all closed $\mathcal{L}_{\mathsf{F}}$-terms t (inclusive of those containing the least number operator).[10] I will return in §5.1 and §5.6 to the comparison of these treatments with latter day interests in axiomatic approaches to truth.

Although Hilbert & Bernays (1939) ultimately group these paradoxes together as *semantische*, the reaction which they register to them varies in detail across the sources we have been considering. What is most important for present purposes is that their discussion of the Liar in (1939, §5.1a) not only immediately precedes their presentation of Gödel's first incompleteness theorem in §5.1c but is also employed as a template for its exposition.[11] This begins as follows:

---

[9] Although Hilbert & Bernays's characterization of "rekursive Zahlentheorie" itself evolved over the two volumes of the *Grundlagen*, it is reasonable to assume at this point in the text that it is a conservative extension by definitions of the system they call Z_μ, which is itself a conservative extension of PA.

[10] Hilbert & Bernays denote this function with the symbol $\mathfrak{e}(x)$ rather than $\mathfrak{d}(x)$. Subsequent authors (e.g. Priest, 1997a; Read, 2016) have tended to view the inconsistency resulting from the assumption that such a function is definable as a *sui generis* antinomy – the so-called *denotational paradox*. But as Hilbert & Bernays themselves observe, the underlying argument is similar to their prior demonstration (1934, p. 330/335) that the denotation function for primitive recursive terms cannot itself be primitive recursive. This goes by a (now) familiar variant of Cantor's diagonal method involving the construction of the "anti-diagonal" function $\mathfrak{a}(x) = \mathrm{subst}(\ulcorner\mathfrak{d}(x)\urcorner, x) + 1$ (which Hilbert & Bernays denote by using a two-place function $\chi(a, n) = \varphi_n(a)$ where $\varphi_0(x), \varphi_1(x), \ldots$ is assumed to be an enumeration of primitive recursive definitions). Wang (1955) later showed how this result can be transformed into an incompleteness theorem by observing that the denotation function for first order-arithmetical terms (including the μ-operator) can be formalized in second-order arithmetic and then invoking the Arithmetized Completeness Theorem in the manner discussed in §4.3 below. The same pattern also arises for the formalization of Berry's paradox — i.e. when the denotation function is formalized either by using a provability predicate (as was carried out by Boolos, 1989) or by the Arithmetized Completeness Theorem (as was carried out by Vopěnka, 1966 and Kikuchi et al., 2012), a similar incompleteness result can be obtained.

[11] See (Sieg & Ravaglia, 2005) for a more complete reconstruction of (1939, §5.1-2).

> From the antinomy of the Liar, we are led to Gödel's theorems ... by considering a modification of that antinomy, and applying the method of formal sharpening. (1939, pp. 269/278)

In §5.1c Hilbert & Bernays provide a detailed exposition of how the proof of the first incompleteness theorem can be formalized in F leading in turn to the second incompleteness theorem. After this they ultimately make the following remark about the significance of the Liar:

> With this demonstration that if F is consistent, the formula C expressing the consistency of F is underivable in F, the formalization of the modified Liar antinomy has reached its full consequence. In fact, this non-derivability is what the ordinary language rendition of the antinomy corresponds to in the sharply circumscribed deductive formalism. (1939, p. 284/294)

This passage provides a precedent for the view which I will suggest in §5 was later developed in a more direct way by Kreisel and Wang —- i.e. that when the everyday notions involved in the paradoxes are analyzed formally and then subjected to the method of arithmetization, the corresponding paradoxical statements are converted into instances of formal incompleteness, thereby dissolving them as potential threats to the development of infinitary mathematics.

Another role played by the paradoxes in the work of Hilbert and his collaborators was as a constraint on the proper development of systems of higher-order logic. In the second edition of the *Grundzüge* (1938), for instance, it is suggested that their discussion of Russell's paradox shows the untenability of an "undifferentiated predicate concept" (p. 120). And in (1917–1918) and (1928) Hilbert suggests that introducing a hierarchy of orders of predicates was sufficient to ensure that analogs of the Liar and the Richard paradoxes could not arise.

On the other hand, the mathematical motivation which is given in these sources for the introduction of higher-order systems is that of formalizing analysis. In (1938) this is carried out in the system which Hilbert and Ackermann state to be equivalent to the simple theory of types up to order ω. But they also observe in §IV.6 that only second-order quantification is required to formalize the existence of least upper bounds for definable sets of real numbers formalized as Dedekind cuts. And in fact a system which we would now recognize as *second-order logic* is explicitly introduced in §IV.1 of the *Grundzüge* to formalize this construction.[12]

The importance of second-order logic to our current topic derives not only from its potential use in axiomatizing analysis, but also from Hilbert & Bernays's recognition that second-order languages allow for the formalization of first-order semantic notions. They first discuss such a possibility in (1939, §5.1a) after remarking that that their axiomatic presentation of the Liar can be regarded as providing an alternative proof of Tarski's Theorem (1935) on the undefinability of truth for formalism F. But at the end of this section they also remark that it is possible to formalize the definition of truth for $\mathcal{L}_{\mathsf{F}}$ in an appropriate higher-order extension of this language.

The possibility of providing a mathematical formalization of the definition of truth for a mathematical language $\mathcal{L}_1$ in an expressively richer language $\mathcal{L}_2$ had, of course, been

---

[12] In (Hilbert & Ackermann, 1928) this is referred to as the *erweiterten Kalkül* but is then rechristened as the *Prädikatenkalkül der zweiten Stufe* in (Hilbert & Ackermann, 1938). A version of this system is adjoined to a first-order arithmetic theory to conduct a more extensive development of analysis in Supplement IV of (Hilbert & Bernays, 1939). These sections are often cited as the origin of the theory now known as *second-order arithmetic* (Z_2) and as an anticipation of Reverse Mathematics (Simpson, 2009). See (Dean & Walsh, 2017, §3) for further discussion of these claims.

foreseen by Gödel (1931b) and is also alluded to repeatedly by Tarski (1935).[13] However, neither Gödel nor Tarski produced such a definition explicitly. On the other hand, in (1939, §5.2e) Hilbert & Bernays not only observed that it suffices to take $\mathcal{L}_2$ to be the language of second-order arithmetic in the case that $\mathcal{L}_1$ is a first-order arithmetical language, but they also explicitly constructed a second-order formula which serves as a truth definition for $\mathcal{L}_{\mathsf{F}}$.

I will return to the details of this definition in §4.3 and to its significance in regard to the foundational project of the *Grundlagen* in §5.1. But these developments also served as the direct antecedent to Kreisel and Wang's work on the relation of the paradoxes to the incompleteness phenomena to which we now turn.

§3. Mathematical preliminaries

3.1. Languages and theories The incompleteness results which will be presented in §4 are most naturally understood in relation to a first-order arithmetical theory Z and a second-order theory S of sets and classes. The respective roles of these theories is as follows: 1) Z must be sufficiently expressive to formalize syntax and sufficiently strong to derive a formalized version of the Arithmetized Completeness Theorem; 2) S must be sufficiently strong to interpret Z and sufficiently expressive to allow for a natural formalization of the set theoretic and semantic notions which figure in the paradoxes.[14]

In their original presentations, Kreisel and Wang took Z to be the system Z_μ of Hilbert & Bernays (1939), corresponding to a conservative extension of first-order Peano arithmetic [PA] with an axiomatization of the least-number operator. But for present purposes it will be convenient to assume that Z is a subsystem of PA itself. Recall that this theory is formulated in the first-order language $\mathcal{L}_{\mathsf{Z}} = \{0, s, +, \times, <\}$ and consists of the base theory PA^- together with the induction schema $\mathrm{Ind}(\Sigma_n)$ for Σ_n-formulas for all $n \in \mathbb{N}$. We will also have occasion below to consider the fragment $\mathrm{I}\Sigma_2$ consisting of PA^- together with $\mathrm{Ind}(\Sigma_2)$.

The status of the theory S is more complicated. Although in his original presentation, Kreisel (1950) only defined this system implicitly, he later made clear in (1953) that S can be taken to be a fragment of a form of what is now called *Gödel-Bernays set theory*. Systems answering to this name (and its various cognates) went through several stages of development which in turn led to distinct axiomatizations. It will thus be useful to be more precise about the relationship between contemporary presentations of GB and the more familiar first-order theory ZF.

Recall that ZF is a one-sorted first-order theory formulated in the language $\mathcal{L}_{\mathsf{ZF}}$ containing the *set variables* $x, y, z, \ldots$ and $\in$ as its sole non-logical symbol. In his original exposition, Bernays described a two-sort theory which also contained class variables $X, Y, Z, \ldots$ which had separate set and class membership relations $x \in y$ and $x \eta Y$.[15] For present purposes, however, it will be convenient to consider the one-sorted axiomatization given

---

[13] See in particular p. 188 footnote 1, pp. 195–196, and pp. 236–237 in Tarski (1935) and also footnote † in the English translation (Tarski, 1956).

[14] As we will see, S can take the form of either a set theoretic system with class variables such as GB or an arithmetical system with number class variables such as ACA_0. Although I will follow the contemporary convention of referring to these theories as "second-order", it should be kept in mind that both systems are based on predicative restrictions to the comprehension scheme. As such, it is equally natural to regard them as *two-sorted first-order theories* whose models are given by specifying separate first- and second-sort domains together with a definition of the membership relation $\in$ which holds between first- and second-sort objects.

[15] Bernays (1937) states that he originally presented a variant of von Nuemann's earlier axiomatization of set theory in lectures delivered in 1929-1930 in Göttingen. He also

by Mendelson (1997, §4.1) which contains only class variables which is in turn similar to the system presented by Gödel (1940).

The language of $\mathcal{L}_{\mathsf{GB}}$ contains $\in$ as its sole non-logical symbol as well as class variables $X, Y, Z, \ldots$ We define $M(x)$ (i.e. "x is a set") as $\exists Y(x \in Y)$ and introduce *set quantifiers* $\forall x\varphi(x)$ and $\exists x\varphi(x)$ as abbreviations for $\forall X(M(X) \to \varphi(X))$ and $\exists X(M(X) \wedge \varphi(X))$. In order to formulate the axioms of GB, we first introduce the mapping $(\cdot)^+ : \mathcal{L}_{\mathsf{ZF}} \to \mathcal{L}_{\mathsf{GB}}$ such that φ^+ results from replacing the set variables $x, y, z, \ldots$ with the class variables $X, Y, Z, \ldots$ in φ (renaming as necessary to avoid clashes) and restricts quantifiers by $M(\cdot)$. Also consider the map $(\cdot)^- : \mathcal{L}_{\mathsf{GB}} \to \mathcal{L}_{\mathsf{ZF}}$ which is defined only for sentences in which all quantifiers are bound by $M(\cdot)$ and which replaces $X, Y, Z, \ldots$ with $x, y, z, \ldots$ (renaming variables as necessary) and removing the relativization.

Equality between classes in GB is axiomatized by

(E1) $\forall Z(X = Y \leftrightarrow Z \in X \leftrightarrow Z \in Y)$
(E2) $\forall X \forall Y \forall Z(X = Y \to (X \in Z \leftrightarrow Y \in Z)$

GB additionally includes the $(\cdot)^+$ translations of each of the following axioms of ZF: Null Set, Pairing, Sum Set, Powerset, Infinity. GB also includes the following class existence axioms corresponding to the so-called *Gödel operations*:

(B1) $\exists X \forall u \forall v(\langle u, v\rangle \in X \leftrightarrow u \in v)$ ($\in$-relation)
(B2) $\forall X \forall Y \exists Z \forall u(u \in Z \leftrightarrow u \in X \wedge u \in Y)$ (intersection)
(B3) $\forall X \exists Z \forall u(u \in Z \leftrightarrow u \notin X)$ (complement)
(B4) $\forall X \exists Z \forall u(u \in Z \leftrightarrow \exists v(\langle u, v\rangle \in X))$ (domain)
(B5) $\forall X \exists Z \forall u \forall v(\langle u, v\rangle \in Z \leftrightarrow u \in X)$ (projection)
(B6) $\forall X \exists Z \forall u \forall v \forall w(\langle u, v, w\rangle \in Z \leftrightarrow \langle v, w, u\rangle \in X)$
(B7) $\forall X \exists Z \forall u \forall v \forall w(\langle u, v, w\rangle \in Z \leftrightarrow \langle u, w, v\rangle \in X)$

Let $\mathrm{Fun}(X)$ express that the class X is a class of ordered pairs which is additionally a function. GB also includes the following version of the axiom of Replacement

(Replacement) $\forall X(\mathrm{Fun}(X) \to \forall x \exists y \forall z(z \in y \leftrightarrow \exists v(\langle v, z\rangle \in X \wedge v \in x)))$

expressing that the image of a set under a (class) function is a set.

The foregoing principles provide a finite axiomatization of GB. A notable property of this system is that it is still possible to prove all instances of the comprehension scheme for so-called *predicative formulas* — i.e. formulas which may contain bound set variables but do not contain bound class variables — without using Replacement. This is the content of what Bernays originally called the *Class Theorem* (1937, p. 72-77):

THEOREM 3.1 *Let $\varphi(x, Y_1, \ldots, Y_m)$ be a predicative formula of $\mathcal{L}_{\mathsf{GB}}$ with free set and class variables as displayed. Then*

$$\mathsf{GB\text{-}Replacement} \vdash \exists Z \forall x(x \in Z \leftrightarrow \varphi(x, Y_1, \ldots, Y_m))$$

The Class Theorem may also be used to show that GB proves all instances of the ZF Separation scheme — i.e.

(Separation) For all $\varphi(x, y_1, \ldots, y_n) \in \mathcal{L}_{\mathsf{ZF}}$ with free variables displayed,

$$\forall u \exists z \forall x(x \in z \leftrightarrow \varphi(x, y_1, \ldots, y_n) \wedge x \in u)$$

---

described such a system in a letter to Gödel dated 3 May 1931 (Feferman et al., 2003, pp. 105-115) which is then further elaborated in a series of seven papers in the *Journal of Symbolic Logic* published between 1937 and 1954 (reprinted in Müller, 1976). See (Kanamori, 2009) for more on the history of Bernays's axiomatization and its relation to those of Von Neumann (1925) and Gödel (1940).

This may be shown either using Replacement or from the following weaker principle which Bernays referred to as the *Subclass axiom*

(Subclasses) $\forall Y \forall u \exists z \forall x (x \in z \leftrightarrow (x \in Y \wedge x \in u))$

stating that the intersection of a class and a set is a set.

Define GB^- to be the result of replacing the axiom Separation with Subclasses in the axiomatiziation of GB and ZF^- to be ZF without the $\mathcal{L}_{\mathsf{ZF}}$-replacement scheme (i.e. Zermelo set theory). It follows from the above that GB and GB^- respectively extend ZF and ZF^- in the following sense: for all $\varphi \in \mathcal{L}_{\mathsf{ZF}}$, if $\mathsf{ZF} \vdash \varphi$, then $\mathsf{GB} \vdash \varphi^+$ and if $\mathsf{ZF}^- \vdash \varphi$, then $\mathsf{GB}^- \vdash \varphi^+$.[16] A more substantial result is that this translation in fact yields a conservative extension in the sense that if φ is in the domain of $(\cdot)^-$ and $\mathsf{GB} \vdash \varphi$, then $\mathsf{ZF} \vdash \varphi^-$ and if $\mathsf{GB}^- \vdash \varphi$, then $\mathsf{ZF}^- \vdash \varphi^-$.[17]

The axiomatization Kreisel originally employed in formulating his incompleteness results is similar to an extension of GB^- with function symbols intended to denote each of the Gödel operations. Suppose, for instance, that we let $T_1, \ldots, T_7$ be Skolem functions for each of the existentially bound variables in $\mathsf{B}_1, \ldots, \mathsf{B}_7$. Let $\mathcal{L}_{\mathsf{S}}$ be $\mathcal{L}_{\mathsf{S}} = \mathcal{L}_{\mathsf{GB}} \cup \{T_1, \ldots, T_7\}$, let B_i^S be the Skolemization of $1 \leq i \leq 7$ using $T_1, \ldots, T_7$, and let S be the theory consisting of GB^- together with $\mathsf{B}_1^S \ldots \mathsf{B}_7^S$. A standard argument shows that S is a conservative extension of ZF^-. And a simple modification of the proof of the Class Theorem also shows that for each predicative formula $\varphi(x_1, \ldots, x_n, Y_1, \ldots, Y_n)$ of $\mathcal{L}_{\mathsf{GB}}$, there exists an $\mathcal{L}_{\mathsf{S}}$-term $T(Y_1, \ldots, Y_n)$ such that

$$\mathsf{S} \vdash \forall x_1, \ldots, x_n (\langle x_1, \ldots, x_n \rangle \in T(Y_1, \ldots, Y_n) \leftrightarrow \varphi(x_1, \ldots, x_n, Y_1, \ldots, Y_n)) \tag{3.1}$$

In the case where $\varphi(x_1, \ldots, x_n)$ does not contain class parameters, the corresponding $\mathcal{L}_{\mathsf{S}}$-term can thus be thought of as providing a name — which I will refer to as a *class term* — for the set defined by the formula.[18]

It is also natural to ask whether there exists a two-sorted theory of arithmetic which extends PA with class axioms in the same way in which GB extends the one-sorted theory ZF. One such theory is the system of second-order arithmetic Z_2 formulated over the

---

[16] To see that GB^- proves all instances of Separation, consider $\varphi(x) \in \mathcal{L}_{\mathsf{ZF}}$ and let u be a set. Since $\varphi(x)^+$ is a predicative formula, the Class Theorem can be used to obtain a class Z such that $\mathsf{GB}^- \vdash \forall x (x \in Z \leftrightarrow \varphi(x))$. But then the existence of a set z such that $\forall x (x \in z \leftrightarrow \varphi(x) \wedge x \in u)$ follows from Subclasses. To see that Subclasses follows from Separation, suppose that Y is a class and x is set. If $Y = \emptyset$, then $Y \cap x$ is a set by Null Set. Otherwise let $a \in Y$ and define a class function F from x to x by $F(z) = z$ if $z \in Y$ and $F(z) = a$ otherwise. Since this definition is predicative, F exists by the Class Theorem. But now $Y \cap x = \{u : \exists z (F(z) = u) \wedge z \in x\}$ exists by Separation.

[17] The analogous result for GB was first shown by Novak (1950) and Mostowski (1950) via a (now) familiar model expansion argument which relies on a set theoretic formalization of the Arithmetized Completeness Theorem — see, e.g., (Lévy, 1976, p. 190) for a concise formulation. Shoenfield (1954) also showed that this result can also be obtained proof theoretically by using the second-epsilon theorem of (Hilbert & Bernays, 1939).

[18] Suppose, for instance, that we introduce the names $E =_{\mathrm{df}} T_1$ for the membership relation, $\overline{X} =_{\mathrm{df}} T_3(X)$ for the complementation function, and $D(X) =_{\mathrm{df}} T_4(X)$ for the domain function. Then $V =_{\mathrm{df}} D(E)$ names the universe, $\emptyset =_{\mathrm{df}} \overline{V}$ names the empty class, etc. Similar axiomatizations of GB-like systems with (official or definitionally introduced) class terms were common prior to the contemporary adoption of ZF — e.g., Gödel (1940), (Shoenfield, 1954), (Bernays & Fraenkel, 1958) and (Vopěnka & Hájek, 1972). The axiomatization which (Kreisel, 1953, p. 48) ultimately proposed to formulate his incompleteness results still differs from these in that it omits both Infinity and principles such as Replacement or Subclasses which are required to obtain the separation scheme for sets. I will return to discuss the role of these principles with respect to Kreisel's (1950) incompleteness result in note 30 below.

two-sorted language $\mathcal{L}_{\mathsf{Z}}^2 = \mathcal{L}_{\mathsf{Z}} \cup \{\in\}$ and containing both number variables $x, y, z, \ldots$ and number class variables $X, Y, Z, \ldots$ The system Z_2 consists of PA^- together with the induction axiom

$$\forall X(0 \in X \wedge \forall x(x \in X \rightarrow x' \in X) \rightarrow \forall x(x \in X)) \tag{3.2}$$

and the full schema comprehension scheme

$$\exists X \forall x(x \in X \leftrightarrow \varphi(x)) \tag{3.3}$$

where $\varphi(x)$ may contain both number and number class variables and parameters (but not X free). If this scheme is limited to a quantifier class $\mathcal{C}$ — e.g. Π^1_1 or Δ^1_1 — then the system obtained is referred to as $\mathcal{C}$-CA_0. And if comprehension is restricted to formulas without bound number class variables, the resulting system is known as ACA_0 (for *arithmetical comprehension*). The latter theory can be shown to be conservative over PA and we will see below that it provides a natural base theory for formalizing the semantic paradoxes.[19]

3.2. Interpretations Substantial use will also be made below of the notion of an *interpretation* of a theory T_2 over the language $\mathcal{L}_2$ in a theory T_1 over the language $\mathcal{L}_1$. Recall that such an interpretation is a mapping $(\cdot)^{\mathsf{i}} : \mathcal{L}_2 \rightarrow \mathcal{L}_1$ which associates an $\mathcal{L}_1$-formula with each non-logical symbol $\alpha(\vec{x})$ of $\mathcal{L}_2$ an $\mathcal{L}_1$-formula $\psi_{\alpha,\mathsf{i}}(\vec{x})$ of appropriate arity, defining constant symbols via definite descriptions as appropriate. This mapping is extended to $\mathcal{L}_1$ formulas by commuting with propositional connectives and restricting quantifiers to an $\mathcal{L}_1$-formula $\delta_{\mathsf{i}}(x)$ which gives the *domain* of $(\cdot)^{\mathsf{i}}$. We additionally require that $\mathsf{T}_1 \vdash \exists x \delta_{\mathsf{i}}(x)$ – i.e. the domain of $(\cdot)^{\mathsf{i}}$ is non-empty – and that for all $\mathcal{L}_2$-formulas φ, if $\mathsf{T}_2 \vdash \varphi$, then $\mathsf{T}_1 \vdash \varphi^{\mathsf{i}}$ – i.e. T_2 proves the images of all theorems of T_2 under $(\cdot)^{\mathsf{i}}$.[20]

The most important source of interpretations in the sequel will be the Arithmetized Completeness Theorem. Another familiar example is the so-called *ordinal interpretation* $(\cdot)^{\mathfrak{s}}$ of the language of arithmetic in that set theory. To formulate this with respect to $\mathcal{L}_{\mathsf{Z}}$ and $\mathcal{L}_{\mathsf{GB}}$, we assume that the theory of ordinal numbers has been developed in GB leading to definitions of $0^{\mathfrak{s}} = \emptyset$ (i.e. the first ordinal), $x +^{\mathfrak{s}} y$ (ordinal addition), $x \times^{\mathfrak{s}} y$ (ordinal multiplication), $x <^{\mathfrak{s}} y$ (ordinal less than) and ω (the least infinite ordinal). We may then define $(\cdot)^{\mathfrak{s}} : \mathcal{L}_{\mathsf{Z}} \rightarrow \mathcal{L}_{\mathsf{GB}}$ as follows: $0^{\mathfrak{s}} = \emptyset$, $\delta(x)^{\mathfrak{s}} = x \in \omega$, $(x = y)^{\mathfrak{s}} = (x = y)$, $(x + y)^{\mathfrak{s}} = x +^{\mathfrak{s}} y$, $(x \times y)^{\mathfrak{s}} = x \times^{\mathfrak{s}} y$, and $(x < y)^{\mathfrak{s}} = x <^{\mathfrak{s}} y$. It is well known that $(\cdot)^{\mathfrak{s}}$ provides an interpretation of PA (and thus also of weaker arithmetical theories) not only in ZF but also in weaker theories like ZF^- and ZF^- - $\mathsf{Infinity}$ – e.g. (Kaye & Wong, 2007). By the results summarized above, this extends to GB^- and related theories such as S.

There is a related interpretation $(\cdot)^{\mathfrak{s}^2}$ of Z_2 and its subsystems in GB^- and thus also in S. In this case, we must revise our original definition of of interpretation to provide definitions $\delta_1(x)$ and $\delta_2(x)$ of its number and number class domains. To this end, we may define $\delta_1(x) = x \in \omega$ as before and $\delta_2(X) = X \in \mathcal{P}(\omega)$. If we now add the inductive clauses $(x \in Y)^{\mathfrak{s}^2} = \delta_1(x) \wedge \delta_2(Y) \wedge x \in Y$, $(\forall X \varphi(X))^{\mathfrak{s}^2} = \forall X(\delta_2(X) \rightarrow \varphi(X)^{\mathfrak{s}^2})$, and $(\exists X \varphi(X))^{\mathfrak{s}^2} = \exists X(\delta_2(X) \wedge \varphi(X)^{\mathfrak{s}^2})$, then it is straightfoward to see that $(\cdot)^{\mathfrak{s}^2}$ provides an interpretation of Z_2 in GB^-.

3.3. The Arithmetized Completeness Theorem Bernays originally proved the Arithmetized Completeness Theorem in the form given by Theorem 2.1. After this,

---

[19] See (Simpson, 2009) for a canonical formulation of these theories.

[20] This is the notion of *relative interpretability* first explicitly introduced in (Tarski et al., 1953, §I.1.5). See (Kreisel, 1950, 1952, 1955) and (Feferman, 1960) for contemporaneous discussions of the origin of this notion in relation to Hilbert and Bernays's method of arithmetization and the Arithmetized Completeness Theorem. For a more complete statement of the standard definitions see, e.g., (Hájek & Pudlák, 1998, pp. 148-150).

a number of generalizations and refinements have been obtained. The first of these was provided by Kreisel (1950) in the course of presenting the first of the incompleteness results which I will formulate in §4. Kreisel observed that not only does the irrefutability of a formula φ in the predicate calculus entail the existence of an arithmetical interpretation $(\cdot)^*$ in which φ^* is true, but that this fact can be formalized in Hilbert & Bernays's theory Z_μ using the notion of interpretability.

To state Kreisel's result in contemporary terms, let $\mathrm{Prov}_{\mathsf{T}}(x)$ be the canonically defined provability predicate for a recursively axiomatizable theory T and $\mathrm{Con}(\mathsf{T})$ its canonical consistency statement $\neg\mathrm{Prov}_{\mathsf{T}}(\ulcorner 0 = 1 \urcorner)$.

THEOREM 3.2 *Let T be a recursively axiomatizable theory over the language $\mathcal{L}_{\mathsf{T}} = \{P_1, \ldots, P_n, f_1, \ldots, f_m, c_1, \ldots, c_k\}$ and let Z extend $\mathrm{I}\Sigma_2$. Then there are Δ^0_2 $\mathcal{L}_{\mathsf{Z}}$-predicates $\delta(x), P_1^\mathfrak{z}, \ldots, P_n^\mathfrak{z}$, $f_1^\mathfrak{z}, \ldots, f_m^\mathfrak{z}$, $c_1^\mathfrak{z}, \ldots, c_q^\mathfrak{z}$ of the appropriate arities such that*

$$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \mathrm{Prov}_{\mathsf{T}}(\ulcorner \varphi \urcorner) \to \varphi^\mathfrak{z} \tag{3.4}$$

where $\varphi^\mathfrak{z}$ is the result of replacing $P_i^\mathfrak{z}$ for P_i $f_i^\mathfrak{z}$ for f_i and $c_i^\mathfrak{z}$ for c_i and relativizing quantifiers with $\delta(x)$. Moreover, $(\cdot)^\mathfrak{z} : \mathcal{L}_{\mathsf{T}} \to \mathcal{L}_{\mathsf{Z}}$ is an interpretation of T in $\mathsf{Z} + \mathrm{Con}(\mathsf{T})$ – i.e. for all $\varphi \in \mathcal{L}_{\mathsf{T}}$, if $\mathsf{T} \vdash \varphi$, then $\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \varphi^\mathfrak{z}$.

In Kreisel's original proof of Theorem 3.2 he constructed the definition of $(\cdot)^\mathfrak{z}$ by showing how it was possible to directly formalize Hilbert & Bernays's (1939) original proof of Theorem 2.1, which in turn follows the method of Gödel (1930).[21] On the other hand, Feferman (1960, pp. 61-62) later realized that it is also possible to formalize Henkin's (1949) more familiar completeness proof in a sufficiently strong fragment of PA. It will be useful to summarize the main steps of this construction so as to illustrate some additional features of the interpretation $(\cdot)^\mathfrak{k}$ which is obtained.[22]

We begin by observing that if T is recursively axiomatizable, then it is possible to obtain a primitive recursive enumeration of all $\mathcal{L}_{\mathsf{T}}$-formulas with the free variable x as $\varphi_1(x), \varphi(x_2), \ldots$ and also an associated countable set of $C = \{c_{\varphi_0}, c_{\varphi_1}, \ldots\}$ of new Henkin constants. On this basis we define the set of witness axioms for $\mathcal{L}_{\mathsf{T}}$

$$H_{\mathsf{T}} = \{\exists x \varphi_i(x) \to \varphi_i(c_i) \ : \ i \in \mathbb{N}\} \tag{3.5}$$

A familiar argument shows that $\mathsf{T} + H_{\mathsf{T}}$ is conservative over T, from which it follows that if T is consistent, then so is $\mathsf{T} + H_{\mathsf{T}}$. We now wish to describe in a manner which can be expressed in $\mathcal{L}_{\mathsf{Z}}$ a method for constructing a *Henkin-complete* extension of T – i.e. a deductively closed set Γ_{T} of $\mathcal{L}^H_{\mathsf{T}} = \mathcal{L}_{\mathsf{T}} \cup C$-sentences extending T such that i) for all formulas φ of this language, either $\varphi \in \Gamma_{\mathsf{T}}$ or $\neg\varphi \in \Gamma_{\mathsf{T}}$, ii) if $\exists x \varphi(x) \in \Gamma_{\mathsf{T}}$, then there is a constant $c \in C$ such that $\varphi(c) \in \Gamma_{\mathsf{T}}$, and iii) if T is consistent, then Γ_{T} is consistent.

As I will discuss further in §5.5, there is by no means a unique method for defining Γ_{T}, even relative to a fixed recursive presentation of T. For present purposes, however, it will be convenient to employ the following familiar construction: i) let $\psi_0, \psi_1, \ldots$ be a recursive enumeration of all $\mathcal{L}^H_{\mathsf{T}}$-sentences; ii) define $\Gamma^0_{\mathsf{T}} = \mathsf{T} + H_{\mathsf{T}}$; iii) $\Gamma^{i+1}_{\mathsf{T}}$ is defined inductively according to the following *extension condition*:

$$\Gamma^{i+1}_{\mathsf{T}} = \begin{cases} \Gamma^i_{\mathsf{T}} \cup \{\psi_i\} & \text{if } \Gamma^i_{\mathsf{T}} \nvdash \neg\psi_i \\ \Gamma^i_{\mathsf{T}} & \text{otherwise} \end{cases} \tag{3.6}$$

We now wish to use the set $\Gamma_{\mathsf{T}} = \bigcup_{i \in \mathbb{N}} \Gamma^i_{\mathsf{T}}$ to define a model of T whose domain comprises the constant symbols C. In so doing we face the familiar problem that since Γ_{T} will contain

[21] The statement appears on p. 266 of (Kreisel, 1950) and the proof is given on pp. 267-273.

[22] See, e.g., (Lindström, 1997, §6.1) for additional details.

sentences of the form $c_{\varphi_i} = c_{\varphi_j}$ (for $i \neq j$), we cannot take the interpretation of c_{φ_i} to be this symbol itself. Since we ultimately wish to define a model arithmetically, one means of circumventing this problem is to extend the Gödel numbering $\ulcorner \cdot \urcorner$ — which we assume is already in place for $\mathcal{L}_\mathsf{T}$ – to C. We can then define the function $h(\ulcorner c_{\varphi_i} \urcorner) = \mu x.(x = \ulcorner c_{\varphi_j} \urcorner \wedge c_{\varphi_i} = c_{\varphi_j} \in \Gamma_\mathsf{T})$ which enumerates the least members of equivalence classes of Henkin constants up to provable equality in Γ_T.

In the general case, $h(x)$ may not be computable. But it is straightforward to see that it is possible to construct a $\mathcal{L}_\mathsf{Z}$-formula $\chi(x)$ defining the function *the Gödel number of the xth constant in the range of $h(x)$* relative to the definition of Γ_T given above. The definition of Γ^{i+1}_T in terms of Γ^i_T can be formalized by a $\mathcal{L}_\mathsf{Z}$-formula $\forall z \eta(x, y, z)$ (for $\eta(x, y, z) \in \Delta^0_1$) expressing that the formula with Gödel number x is a member of Γ^y_T. We can then obtain a Σ^0_2-formula $\gamma(x) =_{\mathrm{df}} \exists y \forall z \eta(x, y, z)$ expressing that there is a stage in the construction such that the formula with Gödel number x is adjoined to Γ_T. Σ^0_2-induction hence suffices to show that the construction of Γ^i_T is defined for all i.

Suppose we also introduce the following standard abbreviations for $\mathcal{L}_\mathsf{Z}$-formulas defining the following notions: $\mathrm{Sent}_{\mathcal{L}_\mathsf{T}}(x)$ for formula expressing the set of Gödel numbers of $\mathcal{L}_\mathsf{T}$-sentences, $\mathrm{Form}^{\dot{v}}_{\mathcal{L}_\mathsf{T}}(x)$ for the formula expressing the set of Gödel numbers of $\mathcal{L}_\mathsf{T}$-formulas with free variable v, $\mathrm{Var}(x)$ for the set of Gödel numbers of variables, $\mathrm{subst}(x, y)$ for the Gödel number of the formula formed by substituting the term with Gödel number y into the first free variable of the formula with Gödel number x, $\mathrm{Prov}_{\mathrm{FOL}}(x)$ for provability of the formula with Gödel number x in pure first-order logic, $\mathrm{Con}(\gamma(x))$ for the consistency of any finite set of formulas satisfying $\gamma(x)$, and $x \cdot y$ for the code of the sequence formed by concatenating the codes x and y.

Making use of the familiar "dot" notation of Feferman (1960), we can now record several facts which collectively formalize that $\gamma(x)$ defines a Henkin-complete extension of T:

(3.7a) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{T}}(x) \wedge \mathrm{Prov}_\mathsf{T}(x) \rightarrow \gamma(x))$$

(3.7b) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \mathrm{Con}(\gamma(x))$$

(3.7c) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \forall x \forall y((\mathrm{Sent}_{\mathcal{L}_\mathsf{T}}(x) \wedge \mathrm{Sent}_{\mathcal{L}_\mathsf{T}}(y) \wedge \gamma(x) \wedge \mathrm{Prov}_{\mathrm{FOL}}(x \dot{\rightarrow} y)) \rightarrow \gamma(y))$$

(3.7d)
$$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \forall x \forall y(\mathrm{Form}^{\dot{y}}_{\mathcal{L}_\mathsf{T}}(x) \wedge \mathrm{Var}(y) \wedge \gamma(\ulcorner \exists \urcorner \cdot y \cdot x) \leftrightarrow \exists u \exists v(\chi(u) = v \wedge \gamma(\mathrm{subst}(x, v))))$$

(3.7e) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{T}}(x) \rightarrow (\gamma(\dot{\neg} x) \leftrightarrow \neg \gamma(x)))$$

These properties respectively express the fact that $\gamma(x)$ defines a set of formulas which extends T, is consistent, is closed under deductive consequence, is closed under the Henkin witness property, and is complete in the sense that it contains the negation of φ just in case it fails to contain φ. It follows from (3.7c,d,e) and the standard interdefinabilties of the connectives and quantifiers that $\gamma(x)$ satisfies compositional clauses similar to those appearing in Tarski's inductive definition of truth (a point to which I will return in §5.5). Note finally that it follows from (3.7e) that a necessary and sufficient condition for the membership of a formula φ in the set Γ_T defined by $\gamma(x)$ is that $\neg\varphi$ is not a member of Γ_T. It follows that we could also define $\gamma(x)$ by the Π^0_2-formula $\forall y \exists z \neg \eta(\dot{\neg} x, y, z)$ and thus that $\gamma(x)$ is in the formula class Δ^0_2.

Using the formulas $\chi(x)$ and $\gamma(x)$ we can now define the interpretation $(\cdot)^\mathfrak{z} : \mathcal{L}_\mathsf{T} \to \mathcal{L}_\mathsf{Z}$ mentioned in Theorem 3.2:

(3.8a) $\quad \delta(x) = \exists u(\chi(u) = x)$

(3.8b) $\quad R(x_1, \ldots, x_n)^\mathfrak{z} = \exists u_1 \ldots \exists u_n(x_1 = \chi(u_1) \wedge \ldots \wedge x_n = \chi(u_n) \wedge \gamma(\ulcorner R(\chi(\dot{u}_1), \ldots, \chi(\dot{u}_n))\urcorner))$

(3.8c) $\quad (x = c)^\mathfrak{z} = \exists u(\chi(u) = x \wedge \gamma(\ulcorner c = \chi(\dot{u})\urcorner)$

(3.8d) $\quad (f(x_1, \ldots, x_n) = y)^\mathfrak{z} = \exists u_1 \ldots \exists u_n \exists u_{n+1}(x_1 = \chi(u_1) \wedge \ldots x_n = \chi(u_n) \wedge \chi(u_{n+1}) = y \wedge \gamma(\ulcorner f(\chi(\dot{u}_1)), \ldots, \chi(\dot{u}_n)) = \chi(\dot{u}_{n+1})\urcorner)$

By an external induction using (3.7a–e) it may now be shown that for all $\mathcal{L}_\mathsf{T}$-formulas $\varphi(x_1, \ldots, x_n)$ with free variables displayed

(3.9) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash (\delta(x_1) \wedge \ldots \wedge \delta(x_n)) \to (\varphi(x_1, \ldots, x_n))^\mathfrak{z} \leftrightarrow (\exists u_1 \ldots \exists u_n(x_1 = \chi(u_1) \wedge \ldots \wedge x_n = \chi(u_n) \wedge \gamma(\ulcorner \varphi(\chi(\dot{u}_1), \ldots, \chi(\dot{u}_n))\urcorner)))$$

Specializing to the case of sentences we thus have

(3.10) $$\mathsf{Z} + \mathrm{Con}(\mathsf{T}) \vdash \varphi^\mathfrak{z} \leftrightarrow \gamma(\ulcorner \varphi \urcorner)$$

for all $\varphi \in \mathcal{L}_\mathsf{T}$.

The foregoing observations comprise what might be called the *syntactic* form of the Arithmetized Completeness Theorem. This version will be sufficient for most of the applications considered below. However it should also be evident that (3.9) can be understood as expressing that $\gamma(x)$ defines truth in an arithmetical model of T described by the interpretation $(\cdot)^\mathfrak{z}$. This point can be appreciated most vividly when T is a $\mathcal{L}_\mathsf{Z}$-theory and the Theorem 3.2 is applied to applied to T itself. For consider a model $\mathcal{M}_1$ of T in which $\mathrm{Con}(\mathsf{T})$ is also satisfied – i.e. $\mathcal{M}_1 \models \mathsf{T} + \mathrm{Con}(\mathsf{T})$. In this case it is also possible to define another model $\mathcal{M}_2 \models \mathsf{T}$ relative to $\mathcal{M}_1$ such that

(3.11) $$\mathcal{M}_1 \models \gamma(\ulcorner \varphi \urcorner) \text{ if and only if } \mathcal{M}_2 \models \varphi$$

by interpreting the non-logical terms of T according to formulas given by $(\cdot)^\mathfrak{k}$ relative to $\mathcal{M}_1$. This can be compared to the standard construction of a model of T from the maximally consistent set Γ_T defined in the non-formalized version of the Henkin construction.

The foregoing observation can also be generalized by modifying the definition of the unary predicate $\gamma(x)$ to obtain a binary predicate $\sigma(x, y)$ which defines satisfaction rather than truth in the following sense.

Definition 3.3 *Let* Z *be an arithmetical theory extending* $\mathrm{I}\Sigma_2$, $\mathcal{M}_1 \models \mathsf{Z}$, *and* $\mathcal{M}_2$ *a structure for the language of a recursively axiomatizable theory* T. *We say that* $\mathcal{M}_2$ *is* strongly definable *in* $\mathcal{M}_1$ *just in case*

i) $|\mathcal{M}_1| = |\mathcal{M}_2|$

ii) *There is an* $\mathcal{L}_\mathsf{Z}$*-formula* $\sigma(x, y)$ *and such that for all* $a_1, \ldots, a_n \in |\mathcal{M}_2|$ *such that* $\varphi(\vec{x}) \in Form_{\mathcal{L}_\mathsf{T}}$, *then*

$$\mathcal{M}_2 \models \varphi(\overline{a}_1, \ldots, \overline{a}_n) \text{ if and only if } \mathcal{M}_1 \models \sigma(\ulcorner \varphi(x_1, \ldots, x_n)\urcorner, \langle \overline{a}_1, \ldots, \overline{a}_n \rangle)$$

where $\langle \ldots \rangle$ *is a coding function for finite sequences.*

The following *semantic* form of the Arithmetized Completeness Theorem can now be obtained (cf., e.g., Smorynski, 1984):

THEOREM 3.4 *Let* $\mathsf{Z} \supseteq \mathrm{I}\Sigma_2$ *and let* T *be a recursively axiomatizable* $\mathcal{L}_\mathsf{T}$*-theory. Then if* $\mathcal{M}_1 \models \mathsf{Z} + \mathrm{Con}(\mathsf{T})$, *there is a model* $\mathcal{M}_2 \models \mathsf{T}$ *which is strongly definable in* $\mathcal{M}_1$. *Moreover, if* T *possesses an infinite model, then the interpretation of the equality symbol given by* $\sigma(x, y)$ *can be the identity relation on* $\mathcal{M}_1$.

In the case where T is itself an $\mathcal{L}_\mathsf{Z}$-theory, another consequence of strong definability is the following:

PROPOSITION 3.5 *Let T be a recursively axiomatizable $\mathcal{L}_\mathsf{Z}$-theory extending PA^-, $\mathcal{M}_1 \models \mathrm{I}\Sigma_2$ be such that $\mathcal{M}_1 \models \mathrm{Con}(\mathsf{T})$ and $\mathcal{M}_2 \models \mathsf{T}$ be strongly definable in $\mathcal{M}_1$. Then there is an $\mathcal{M}_1$-definable embedding $f : |\mathcal{M}_1| \to |\mathcal{M}_2|$ onto an initial segment of $\mathcal{M}_2$.*

Underlying this result is the observation that if $\mathcal{M}_2$ is strongly definable in $\mathcal{M}_1$ then there exist $\mathcal{L}_\mathsf{Z}$-formulas $\psi_0(x)$ and $\psi_s(x, y)$ which define the least element element $0^{\mathcal{M}_2}$ and successor function $s^{\mathcal{M}_2}(x)$ of $\mathcal{M}_2$ relative to $\mathcal{M}_1$. Letting $(t)_i$ denote a standard coding of finite sequences, the graph of the function $f(x)$ can thus be defined by the formula

$$\psi_f(x, y) = \exists t(\mathrm{len}(t) = x + 1 \wedge \psi_0((t)_0) \wedge \forall i < x(\psi_s((t)_i, (t)_{i+1})) \wedge (t)_x = y) \tag{3.12}$$

As $\psi_f(x, y)$ is Σ^0_1 in $\psi_0(x)$ and $\psi_s(x, y)$, it can then be verified by Σ^0_2-induction that $f(x)$ preserves sums and products, is one-one and onto an initial segment. It thus follows that if $\mathcal{M}_2 \models \mathsf{T}$ is strongly definable in $\mathcal{M}_1$ then $\mathcal{M}_2$ is an *end extension* of $\mathcal{M}_2$.[23]

Another important consequence of the definability of the embedding $f(x)$ is that it provides a means of formalizing substitution into the image of a formula under an interpretation obtained from the Arithmetized Completeness Theorem. Suppose, for instance, that $(\cdot)^* : \mathcal{L}_\mathsf{T} \to \mathcal{L}_\mathsf{Z}$ is an interpretation of an $\mathcal{L}_\mathsf{Z}$-theory T in $\mathsf{Z} + \mathrm{Con}(\mathsf{T})$ as above. We also introduce the symbol $\mathrm{f}(x)$ to abbreviate the function defined by $\psi_f(x, y)$ and write $\varphi^*(x)$ to denote the result of applying the inductive clauses in the definition of $(\cdot)^*$ to $\varphi(x)$ while treating x as a free variable. $\varphi^*(\mathrm{f}(\overline{n}))$ thus expresses that $\varphi(x)$ holds of the nth element of the model defined by $(\cdot)^*$. More generally, the following may now be shown by external induction on n:

COROLLARY 3.6 *For all $\mathcal{L}_\mathsf{Z}$-formulas $\varphi(x)$ and $n \in \mathbb{N}$,*

$$\mathsf{Z} \vdash (\varphi(\overline{n}))^* \leftrightarrow \varphi^*(\mathrm{f}(\overline{n})) \tag{3.13}$$

As we will see in the next section, the definability of $f(x)$ is one of several aspects of the Arithmetized Completeness Theorem which allows us to obtain statements which provably differ in truth value between the models $\mathcal{M}_1$ and $\mathcal{M}_2$ obtained in Theorem 3.5 and which thereby must be independent of T.

3.4. Arithmetized completeness for* GB *and related systems Applying Theorem 2.1 to a finitely axiomatizable $\mathcal{L}_\mathsf{S}$-theory such as GB or S yields an arithmetical formula $(x \in y)^\mathfrak{z}$ which interprets the membership relation relative to Z. By next applying Theorem 3.4 in the case where $\mathcal{M}_1$ is the standard model of arithmetic $\mathcal{N} = \langle \mathbb{N}, 0, s, +, \times, < \rangle$ we may also obtain a model either of the form $\mathcal{M}_2 = \langle \mathbb{N}, E \rangle$ for GB or of the form $\mathcal{M}_2 = \langle \mathbb{N}, E, B_1, \ldots, B_7 \rangle$ for S where $E \subseteq \mathbb{N} \times \mathbb{N}$ is the extension of $(x \in y)^\mathfrak{z}$ in $\mathcal{N}$ and $B_i : \mathbb{N}^j \to \mathbb{N}$ are interpretations $T_i^\mathfrak{z}$ of the Skolem terms T_i corresponding to the Gödel operations.

Theorem 3.2 highlights how an arithmetical system such as $\mathrm{I}\Sigma_2$ is a natural theory in which to formalize this construction. But it is also clear that we could work over any theory T for which there is an interpretation $(\cdot)^\mathfrak{t} : \mathcal{L}_\mathsf{Z} \to \mathcal{L}_\mathsf{T}$ of $\mathrm{I}\Sigma_2$ in T. In this case, Theorem 3.2 also provides an interpretation $(\cdot)^\mathfrak{z}$ of T itelf over $\mathrm{I}\Sigma_2 + \mathrm{Con}(\mathsf{T})$. But since

[23] A version of Proposition 3.5 is assumed at one point in Kreisel's (1950, p. 273) original incompleteness proof. However it appears to be Wang (1955, p. 36) who first formulated the definition of $f(x)$ explicitly. See (Smorynski, 1984, p. 41) for references to later rediscoveries and applications of Proposition 3.5.

the composition of interpretations is itself an interpretation, $((\cdot)^{\mathfrak{z}})^{\mathfrak{t}} : \mathcal{L}_{\mathsf{T}} \to \mathcal{L}_{\mathsf{T}}$ also provides an interpretation of T in $\mathsf{T} + \mathrm{Con}(\mathsf{T})$.[24]

In the first case we will be interested in below we will have $\mathsf{T} = \mathsf{S}$. Since all models of S are infinite, Theorem 3.4 allows us to obtain an interpretation $(\cdot)^{\mathfrak{z}} : \mathcal{L}_{\mathsf{S}} \to \mathcal{L}_{\mathsf{Z}}$ for which $\delta_{\mathfrak{z}}(x)$ is $x = x$ — i.e. the domain of $(\cdot)^{\mathfrak{z}}$ is taken to be the set of all natural numbers. Now define $(\cdot)^* : \mathcal{L}_{\mathsf{S}} \to \mathcal{L}_{\mathsf{S}}$ to be the result of composing $(\cdot)^{\mathfrak{z}}$ with the ordinal interpretation $(\cdot)^{\mathfrak{s}} : \mathcal{L}_{\mathsf{Z}} \to \mathcal{L}_{\mathsf{S}}$. Note that in this case we will have $\delta_*(x) = x \in \omega$, from which it follows that all quantifiers in the image φ^* of an $\mathcal{L}_{\mathsf{S}}$-formula φ under the composite interpretation will be relativized to ω.

Note also that $(\cdot)^*$ respectively associates an $\mathcal{L}_{\mathsf{S}}$-formula $\psi_{T_1}(x, y), \psi_{T_2}(x_1, x_2, y), \ldots, \psi_{T_7}(x, y)$ with the terms $T_1, \ldots, T_7$ denoting the Gödel operations such that the former are provably equivalent to arithmetical formulas in S. In the arithmetical case, the analogous observation allows us to obtain an arithmetical formula defining the embedding function introduced in Proposition 3.5. This is possible in part because Z proves the statement $\forall x \exists t((t)_0 = 0 \wedge \forall i < x((t)_{i+1} = (t)_i + 1) \wedge x = (t)_x)$ expressing that every number is obtained by iterating the successor function and is hence canonically denoted by a numeral. But although no analogous result about sets is assumed with respect to S, it is still possible to formalize substitution of class terms under the scope of $(\cdot)^*$ in S in a manner which parallels Corollary 3.6.

Say that a class Y is *nameable* if there exists some closed class term T such that S proves $Y = T$. It is a consequence of the Class Theorem and (3.1) that all predicatively definable classes are namable. But as the set $\mathrm{ClTerm}_{\mathcal{L}_{\mathsf{S}}}$ of closed class terms is clearly countable, it is also not difficult to construct a formula $\nu(x, Y)$ of $\mathcal{L}_{\mathsf{S}}$ which defines the graph of a bijection between ω and $\mathrm{ClTerm}_{\mathcal{L}_{\mathsf{S}}}$. Following Myhill (1952) we may, for instance, start out by assigning a code number to each of the primitive terms $T_1, \ldots, T_7$ and then define a function $g(x)$ which enumerates the codes of all class terms in some standard way which is recursive according to their structure. $\nu(y, Y)$ can then be taken to assert that y is a finite sequence of codes in the range of $g(x)$ which form a formation sequence such that there is a corresponding finite sequence of classes formed by applying Gödel operations in the analogous way whose final element is Y. Note that this naturally corresponds to a Σ^1_1-formula of $\mathcal{L}_{\mathsf{S}}$.

Suppose we now introduce the abbreviation $\mathrm{N}(y)$ for the function defined by $\nu(y, Y)$. Myhill considered the axiom

$$\forall X \exists y \in \omega(\mathrm{N}(y) = X) \tag{3.14}$$

asserting that all classes are namable and showed that it is consistent with GB together with Gödel's (1940) Axiom of Constructibility (i.e. $V = L$). However, what will be of most use below is his demonstration that over GB, $V = L$ also entails that each non-empty namable class has a nameable member — i.e.

THEOREM 3.7 (Myhill, 1952) $\mathsf{GB} + V = L \vdash \forall X(X \neq \emptyset \wedge \exists y \in \omega(\mathrm{N}(y) = X) \rightarrow \exists x(x \in X \wedge \exists z \in \omega(\mathrm{N}(z) = x)))$

The proof of this result relies on the fact that over $\mathsf{GB} + V = L$ it is possible to define in $\mathcal{L}_{\mathcal{S}}$ a well-ordering of the constructible sets. Following (Gödel, 1940, §V), this can then

[24] In the case where T is not an $\mathcal{L}_{\mathsf{Z}}$-theory $\mathrm{Con}(\mathsf{T})$ must here be understood to denote the interpretation of the canonical $\mathcal{L}_{\mathsf{Z}}$ consistency statement for T in $\mathcal{L}_{\mathsf{T}}$ – i.e. what we would otherwise denote by $(\mathrm{Con}(\mathsf{T}))^{\mathfrak{t}}$. Note also that the *identity interpretation* $(\cdot)^{\mathfrak{i}}$ which translates all symbols of $\mathcal{L}_{\mathsf{T}}$ as themselves is also clearly an interpretation of T in $\mathsf{T} + \mathrm{Con}(\mathsf{T})$. But whereas for this interpretation we have $\mathsf{T} \vdash \varphi \leftrightarrow \varphi^{\mathfrak{i}}$, we will see below that this cannot be the case for all φ for the sorts of interpretations obtained from the Arithmetized Completeness Theorem.

be used to obtain a predicatively defined choice function As which when applied to a non-empty class X returns a "designated" set $As(X) \in X$. It may be shown that namable classes are constructible and also that if X is namable, then so is $As(X)$. Thus since As is definable in terms of the Gödel operations, it can finally be shown that there is a class term As denoting As such that

$$\mathsf{S} + V = L \vdash \forall X((X \neq \emptyset \wedge \exists y \in \omega(\mathrm{N}(y) = X)) \rightarrow \mathrm{As}(X) \in X) \tag{3.15}$$

As we will see in §4.2, while the naming function $\mathrm{N}(y)$ is definable in $\mathcal{L}_{\mathsf{S}}$, it cannot be defined by any *predicative* formula (unless is S is inconsistent). On the other hand, all of the quantifiers in the formulas $\psi_{T_1}(x, y), \psi_{T_2}(x_1, x_2, y), \ldots, \psi_{T_7}(\vec{x}, y)$ interpreting the Gödel operations are relativized to ω. Since these formulas are thus predicative, it follows that S supplies corresponding class terms $T_1^*, \ldots, T_7^*$. Now let $\mathrm{ClTerm}^*_{\mathcal{L}_{\mathsf{S}}}$ be the corresponding set of closed class terms constructed from these class terms as primitives. The previous definition of $\nu(y, Y)$ can now be modified to obtain a predicative formula $\nu^*(y, Y)$ defining a bijection between ω and $\mathrm{ClTerm}^*_{\mathcal{L}_{\mathsf{S}}}$. If we now introduce the abbreviation $\mathrm{N}^*(y)$ for the function defined by this formula, $\mathrm{N}^*(0), \mathrm{N}^*(1), \ldots$ then provides an enumeration of classes denoted by the terms in $\mathrm{ClTerm}^*_{\mathcal{L}_{\mathsf{S}}}$ which in turn corresponds to namable classes in the model determined by the interpretation $(\cdot)^*$.

Now suppose that $T \in \mathrm{ClTerm}_{\mathcal{L}_{\mathsf{S}}}$ and $n \in \mathbb{N}$ are such that $\mathsf{S} \vdash \mathrm{N}(n) = T$ (where n is now understood to abbreviate a term denoting the nth finite ordinal). Note that in this case we will also have that $\mathrm{S} \vdash (\mathrm{N}(n) = T)^*$ and thus also $\mathrm{S} \vdash \mathrm{N}^*(\mathrm{f}(n)) = T^*$.[25] Putting these observations together, we thus have for all $\mathcal{L}_{\mathsf{S}}$-formulas $\varphi(x)$

$$\mathsf{S} \vdash (\varphi(T))^* \leftrightarrow \varphi^*(T^*) \leftrightarrow \varphi^*(\mathrm{N}^*(\mathrm{f}(n))) \tag{3.16}$$

This shows that if we restrict attention to namable classes, then we are indeed able to formalize substitution into the scope of an interpretation of $(\cdot)^*$ of S in $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ in parallel to (3.13).

§4. Paradoxes and incompleteness The goal of this section will be to illustrate how the Arithmetized Completeness Theorem provides a tool for transforming a range of familiar paradoxes into formal incompleteness results. The general method will be as follows: i) it is first shown how a given "paradoxical notion" may be naturally formalized as a predicate $\varphi(x)$ in the language of a second-order theory T which interprets Z; ii) it is then shown that the construction of a statement $\psi_{\varphi(x)}$ obtained from $\varphi(x)$ in the manner of one of the original paradoxes (e.g. a Russell-like or Liar-like sentence) does not yield a contradiction but rather is typically *decidable* over T; iii) Theorem 3.2 is then invoked to obtain an arithmetic interpretation $(\cdot)^*$ of T which is used to reinterpret the non-logical terms appearing in $\psi_{\varphi(x)}$ leading to a statement $(\psi_{\varphi(x)})^*$ which is provably equivalent to an $\mathcal{L}_{\mathsf{Z}}$-formula over T; iv) $\psi_{\varphi(x)}$ is finally shown to be *undecidable* over T relative to an appropriate consistency assumption.

I will return in §5.2 to discuss both the apparent uniformity of this method and its bearing on the paradoxes in their original non-arithmetized forms. But before getting underway, it will be useful to first isolate a result which will often be sufficient to demonstrate the undecidability of statements constructed in the manner just described.

4.1. The undecidability lemma Suppose that T interprets $\mathrm{I}\Sigma_2$ and $(\cdot)^* : \mathcal{L}_{\mathsf{T}} \rightarrow \mathcal{L}_{\mathsf{T}}$ is an interpretation of T in $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ which is obtained in the manner described

[25] Here $\mathrm{f}(x)$ should be understood as abbreviating the image of the definition of the embedding $f(x)$ given in §3.3 under the ordinal interpretation whereas T^* abbreviates the class term obtained by uniformly replacing T_i with T_i^* in T.

in §3.3 from a predicate $\gamma(x)$ describing a Henkin-complete extension of T in $\mathcal{L}_{\mathsf{T}}$.[26] Note that in this case $\gamma(x)$ and $(\cdot)^*$ will satisfy the following properties for all $\mathcal{L}_{\mathsf{T}}$-sentences φ, respectively as a consequence of (3.7a,e) and (3.10):

(4.1a) If $\mathsf{T} \vdash \varphi$, then $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \gamma(\ulcorner\varphi\urcorner)$.

(4.1b) $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \gamma(\ulcorner\neg\varphi\urcorner) \leftrightarrow \neg\gamma(\ulcorner\varphi\urcorner)$

(4.1c) $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \varphi^* \leftrightarrow \gamma(\ulcorner\varphi\urcorner)$

LEMMA 4.1 *Suppose that* T *and* $\gamma(x)$ *are as above and* φ *is any* $\mathcal{L}_{\mathsf{T}}$*-sentence such that* $\mathsf{T} \vdash \varphi \leftrightarrow \neg\gamma(\ulcorner\varphi\urcorner)$*. Then if* $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ *is consistent, then* $\mathsf{T} \nvdash \varphi$ *and* $\mathsf{T} \nvdash \neg\varphi$*.*

Proof. Let φ be such that $\mathsf{T} \vdash \varphi \leftrightarrow \neg\gamma(\ulcorner\varphi\urcorner)$. We now consider the two cases:

i) Suppose that $\mathsf{T} \vdash \varphi$ and thus also that $\mathsf{T} \vdash \neg\gamma(\ulcorner\varphi\urcorner)$. In virtue of the former supposition and (4.1a) we also have that $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \gamma(\ulcorner\varphi\urcorner)$ and thus also $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \bot$.
ii) Suppose that $\mathsf{T} \vdash \neg\varphi$ and thus also that $\mathsf{T} \vdash \gamma(\ulcorner\varphi\urcorner)$. In virtue of the former supposition and (4.1a) we also have that $\mathsf{T} \vdash \gamma(\ulcorner\neg\varphi\urcorner)$. But then by (4.1b) we have $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \neg\gamma(\ulcorner\varphi\urcorner)$ and thus also $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \bot$.

□

As a version of the Diagonal Lemma will be available for the sort of theory in question (either directly or by interpretation), it is possible to directly construct a sentence γ_λ such that $\mathsf{T} \vdash \lambda_\gamma \leftrightarrow \neg\gamma(\ulcorner\lambda_\gamma\urcorner)$. Since the consistency of $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ is entailed by the ω-consistency of T (and in fact even by its 1-consistency), Lemma 4.1 already yields a version of Gödel's First Incompleteness Theorem.[27] Now suppose $\mathcal{M}_1 \models \mathsf{T} + \mathrm{Con}(\mathsf{T})$. Then by Theorem 3.4, there is a model $\mathcal{M}_2$ which is strongly defined in a $\mathcal{M}_1$. It thus also follows from (3.11) that $\mathcal{M}_1 \models \lambda_\gamma$ if and only if $\mathcal{M}_2 \not\models \lambda_\gamma$. It hence follows that $\mathcal{M}_1$ and $\mathcal{M}_2$ cannot be elementarily equivalent – i.e. $\mathcal{M}_1 \not\equiv \mathcal{M}_2$. Taken together with Proposition 3.5, Lemma 4.1 thus also shows that $\mathcal{M}_2$ not only must be an end extension of $\mathcal{M}_1$, but it also must additionally be *proper*. In particular, if $\mathcal{M}_1$ is the standard model $\mathcal{N}$, $\mathcal{M}_2$ must be nonstandard.

Another consequence of Lemma 4.1 which arises from (4.1c) is the following:

COROLLARY 4.2 *Suppose that* T *satisfies the same conditions as in Lemma 4.1 and that* φ *is such that* $\mathsf{T} \vdash \varphi \leftrightarrow \neg\varphi^*$*. Then if* $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ *is consistent, then* $\mathsf{T} \nvdash \varphi$ *and* $\mathsf{T} \nvdash \neg\varphi$

It follows that any sentence which is provably equivalent to the negation of its interpretation under $(\cdot)^*$ must be independent of T as long as $\mathsf{T}+\mathrm{Con}(\mathsf{T})$ is consistent. As we will see, this result is already sufficient to convert many paradoxes which are traditionally described as involving self-reference into formal incompleteness theorems, in some instances without overt diagonalization.

4.2. Set theoretic paradoxes

[26] In the cases where $\mathcal{L}_{\mathsf{T}}$ does not include $\mathcal{L}_{\mathsf{Z}}$, $\gamma(x)$ should be understood in the sequel as the image of the formula constructed in §3.3 in $\mathcal{L}_{\mathsf{Z}}$ under an appropriate interpretation.

[27] A less general form of Lemma 4.1 is presented in this guise by Smorynski (1977, p. 861) whereas Corollary 4.2 (which will be of more use below) is implicitly employed by (Wang, 1955).

4.2.1. Russell's paradox The Class Theorem can be used to illustrate how the traditional set theoretic paradoxes are resolved relative to theories such as GB or S which formally distinguish between sets and classes. For instance, since $\rho(X) = X \notin X$ is a predicative formula, GB proves the existence of the class of sets which do not contain themselves as a member — i.e. $\mathsf{GB} \vdash \exists Y \forall x(x \in Y \leftrightarrow x \notin x)$. It thus follows that there is a term R denoting the *Russell class* such that $\mathsf{S} \vdash \forall X(M(X) \rightarrow (X \in R \leftrightarrow X \notin X))$. From this it follows that $\mathsf{S} \vdash M(R) \rightarrow (R \in R \leftrightarrow R \notin R)$ and hence also $\mathsf{S} \vdash \neg M(R)$. But then $\mathsf{S} \vdash R \notin R$ — i.e. the "Russell sentence" is refutable in S. Although the language of GB does not provide a class term by which this fact can be asserted, it still proves $\neg\exists Y(M(Y) \wedge \forall x(x \in Y \leftrightarrow x \notin x))$. Thus while both GB and S prove that R exists as a class, they embody the traditional resolution to Russell's paradox in the sense that they *refute* that it exists as a set.

If we assume the Axiom of Regularity, then it may also be shown formally in S that no set is a member of itself and thus that $R = V$.[28] Recall, however, that we have seen in §3.3 that it is also possible to uniformly enumerate the namable classes over S via the $\mathcal{L}_{\mathsf{S}}$-definable function $\mathrm{N}(x)$. Although there are only countably many such classes, it is also possible to formulate a version of Russell's paradox for namable classes. For consider the predicate

$$\rho'(x) = x \in \omega \wedge x \notin \mathrm{N}(x) \tag{4.2}$$

which holds of a finite ordinal just in case it is not a member of the class which it names. Now suppose that there existed a term $R' \in \mathrm{ClTerm}_{\mathcal{L}_{\mathsf{S}}}$ such that $\mathsf{S} \vdash \forall X(M(X) \rightarrow (X \in R' \leftrightarrow \rho'(X)))$. In this case we would have that $\mathsf{S} \vdash \mathrm{N}(r') = R'$ for some $r' \in \mathbb{N}$. But then $\mathsf{S} \vdash M(r') \rightarrow (r' \in R \leftrightarrow r' \notin \mathrm{N}(r'))$ and thus also $\mathsf{S} \vdash M(r') \rightarrow (r' \in R' \leftrightarrow r' \notin R')$. It thus follows that if our assumption held, we would also have $\mathsf{S} \vdash M(r')$ (since $r' \in \omega$ which is a set by $\mathsf{Infinity}$) and in fact also $\mathsf{S} \vdash M(R')$ (since R' would be a subclass of ω and hence a set by either $\mathsf{Subclasses}$ or $\mathsf{Replacement}$).

The reductio assumption in this case is thus not that $\rho'(x)$ determines a *set*, but rather that it determines a *class*. For note that since the definition of $\mathrm{N}(x)$ given in §3.4 is Σ^1_1, the definition of $\rho'(x)$ which has been given is not a predicative formula and is thus not guaranteed to form a class denoted by a term of $\mathcal{L}_{\mathsf{S}}$. The argument just given thus demonstrates that as long as S is consistent, there can be no predicative formula which is provably equivalent to $\mathrm{N}(x)$ over S. In particular, we can conclude that GB does not prove that the class function denoted by $\mathrm{N}(x)$ is itself namable, and also that it is consistent with GB that it does not exist as a class.[29]

It is possible to reconstruct Kreisel's original incompleteness result for a system similar to S on the basis of a similar series of observations. To this end, let $(\cdot)^*$ be an interpretation of S in $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ obtained in the manner of §3.3 by composing an arithmetical interpretation $(\cdot)^{\mathfrak{z}}$ of S in $\mathsf{Z} + \mathrm{Con}(\mathsf{S})$ with the ordinal interpretation $(\cdot)^{\mathfrak{k}}$ of Z in S. Now consider the formula

$$\kappa(x) = (x \in \omega \wedge \mathrm{f}^*(x) \notin^* \mathrm{N}^*(\mathrm{f}^*(x))) \tag{4.3}$$

Then $\kappa(x)$ holds of a natural numbers just in case its image in the model determined by $(\cdot)^*$ fails to fall under the class which it names in the sense of that model. Recall that the interpretation $(\cdot)^*$ restricts all quantifiers by the predicate $x \in \omega$ and note also that the definition of the function $\mathrm{N}^*(x)$ given in §3.4 is predicative. Thus unlike $\rho'(x)$, $\kappa(x)$ is a predicative formula of $\mathcal{L}_{\mathsf{S}}$.

---

[28] This appears to have been first explicitly noted by Gödel (1940, p. 10).

[29] For more on this point, see (Kruse, 1963) and §5.2.

From this it follows that for an appropriate class term K,

$$\mathsf{S} \vdash \forall x(x \in K \leftrightarrow \kappa(x)) \tag{4.4}$$

Since K is a fixed member of $\mathrm{ClTerm}_{\mathcal{L}_\mathsf{S}}$, there is also a natural number k such that $\mathsf{S} \vdash \mathrm{N}(k) = K$. But in contradistinction to the situation with the modified Russell class R', we have the following:

PROPOSITION 4.3 *If* $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ *is consistent, then* $\mathsf{S} \nvdash k \in K$ *and* $\mathsf{S} \nvdash k \notin K$.

Proof. From (4.4) we have

$$\mathsf{S} \vdash k \in K \leftrightarrow \mathrm{f}^*(k) \notin^* \mathrm{N}^*(\mathrm{f}^*(k)) \tag{4.5}$$

It follows from (3.13) and (3.16) that the righthand side of this biconditional is equivalent to $(k \in K)^*$. We hence have that

$$\mathsf{S} \vdash k \in K \leftrightarrow \neg(k \in K)^* \tag{4.6}$$

And from this it follows from Corollary 4.2 that if $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ is consistent, then $\mathsf{S} \nvdash k \in K$ and $\mathsf{S} \nvdash k \notin K$. □

Proposition 4.3 represents a partial reconstruction of the original incompleteness result reported by Kreisel (1950, pp. 273-274).[30] Note that the definition of K entails both that it is a set and also that the biconditional (4.6) holds. But rather than leading to the conclusion that $\rho(x)$ does not define a set in the manner of the original paradox or that $\rho'(x)$ does not define a class in the manner of reformulation with namable classes, the biconditional (4.6) entails that S leaves undecided whether the number k is a member of the class which it names. This is our first instance of the pattern described above.

4.2.2. Other set theoretic paradoxes Many familiar set theoretic antinomies are similar to Russell's paradox in that they can be derived by asking whether a certain class is a member of itself. For instance, the paradoxes of Burali-Forti or Mirimanoff can be formulated over a theory such as GB (or S) as follows: i) the relevant set theoretic notions — i.e. *ordinal number* and *well-founded* – can be formalized as $\mathcal{L}_\mathsf{S}$-formulas $Ord(X)$ and $WF(X)$ which respectively hold of the sets which are ordinals or are well-founded ; ii) as these formulas are predicative, they determine classes O and W; iii) under the assumptions $M(O)$ or $M(W)$, it can be shown that the self-membership claims $O \in O$ and $W \in W$ would hold in virtue of the fact that O and W satisfy the definitions of ordinality or well-foundedness; iv) but this can also be shown to contradict the definitions of $Ord(X)$ and $WF(X)$.

On the other hand, a notable difference between these cases and Russell's paradox is that the formulas conventionally chosen for $Ord(X)$ and $WF(X)$ do not instantiate schema which are inconsistent over pure first order logic.[31] In fact, the typical formulations of the Burali-Forti and Mirimanoff paradoxes require us to choose particular formulas defining ordinality or well-foundedness which in turn require the adoption of additional

[30] One obstacle to providing an exact reconstruction of Kreisel's (1950) original result is that he initially failed to provide an axiomatization of the system for which he intended it to hold. Although axiomatization of a system similar to S is given in (Kreisel, 1953), a more serious obstacle is that Kreisel also does not explicitly define the functions $f(x)$ or $N(x)$ which we have seen are required to refer to objects in the model defined by $(\cdot)^*$. It is thus not clear from his proof why K should correspond to a set in the system in which he is operating. For more on Kreisel's original proof, see note 55 below.

[31] Recall in particular that the instance of Comprehension required to derive Russell's paradox has the form $\exists x \forall y(\varphi(x,y) \leftrightarrow \neg\varphi(y,y))$ (taking $x \in y$ for $\varphi(x,y)$), all instances of which are refutable in first-order logic.

set theoretic axioms in order to derive the relevant contradictions.[32] As a consequence, it is less straightforward to transform these paradoxes directly into incompleteness results.

A case of intermediate complexity is provided by Quine's (1981, p. 130) *paradox of reciprocated classes*. Consider the family of predicates $\upsilon_0(x) = x \notin x$ and for $n > 0$

$$\upsilon_n(x) = \neg\exists y_1 \dots y_n(x \in y_1 \wedge y_1 \in y_2 \wedge \dots \wedge y_{n-1} \in y_n \wedge y_n \in x) \tag{4.7}$$

For each n, $\upsilon_n(x)$ holds of those sets x which do not contain an $\in$-cycle of length n. As $\upsilon_i(x)$ is predicative, S proves the existence of the class U_i determined by each predicate $\upsilon_i(x)$. But it is also easy to see that S proves $U_0 = R$ and thus also $\neg M(U_0)$. The fact that $\mathsf{S} \vdash \neg M(U_i)$ for $i \geq 1$ illustrates the general pattern of reasoning required to formally derive $\neg M(O)$ or $\neg M(W)$.[33] By way of example, consider the case for $n = 1$ and assume for reductio that $M(U_1)$. From this it follows that

$$U_1 \in U_1 \leftrightarrow \neg\exists y(U_1 \in y \wedge y \in U_1) \tag{4.8}$$

Reasoning in S, we may now proceed as follows:

i) Suppose $U_1 \in U_1$.
ii) Then $\neg\exists y(U_1 \in y \wedge y \in U_1)$ by (4.8).
iii) From i) we have $U_1 \in U_1 \wedge U_1 \in U_1$. But by generalizing we obtain $\exists y(U_1 \in y \wedge y \in U_1)$ contradicting ii).

iv) Suppose $U_1 \notin U_1$.
v) Then $\exists y(U_1 \in y \wedge y \in U_1)$ by (4.8).
vi) If we fix y such that $U_1 \in y \wedge y \in U_1$ it then follows that $\exists z(z \in y \wedge y \in z)$.
vii) Since $y \in U_1$, $\neg\exists z(z \in y \wedge y \in z)$ by (4.8) again. But this contradicts vi).

Since $\mathsf{S} \vdash U_1 \in U_1 \vee U_1 \notin U_1$, i)–vii) comprise a derivation of $\neg M(U_1)$ in S.

As an intermediate step in developing an arithmetized version of this paradox, note that it is again possible to formulate a version of this reasoning for the namable classes analogous to the second version of the Russell antinomy presented above. Consider in particular the predicate

$$\upsilon_1'(x) = x \in \omega \wedge \neg\exists y(x \in \mathrm{N}(y) \wedge y \in \mathrm{N}(x)) \tag{4.9}$$

which is obtained from $\upsilon_1(x)$ in the same manner by which $\rho'(x)$ is obtained from $\rho(x)$ above. If we were to assume for reductio that $\upsilon_1'(x)$ determines a class U_1', then it would follow that there exists a natural number $u_1' \in \mathbb{N}$ such $\mathsf{S} \vdash \mathrm{N}(u_1') = U_2'$ for an appropriate class term U_1'. Thus since $\mathsf{S} \vdash U_1' \subseteq \omega$, we would again have $M(U_1')$ by $\mathsf{Subclasses}$. But as in the case of $\rho'(x)$ it is easy to see that a contradiction can be reached by mimicking the reasoning of the argument i)–vii).

In order to obtain an incompleteness result analogous to Proposition 4.3, we now wish to replicate the forgoing reasoning under the scope of an interpretation obtained from the Arithmetized Completeness Theorem. But we will see that this requires that we are

[32] See (Doets, 1999) for a discussion of the additional principles which are required in these cases.

[33] In fact the paradox of reciprocated classes can be understood as a simplification of Montague's (1955) *paradox of grounded classes* which in turn can be understood as a simplification of Mirimanoff paradoxes which avoids the need to formalize the notion of finitude. Montague's version arises when we consider the class of all *grounded* sets satisfying the predicate $\alpha(x) = \forall u(x \in u \rightarrow \exists y(y \in u \wedge \neg\exists z(z \in u \wedge z \in y))$ which holds of x just in case all sets u containing x have the property which is asserted to hold of all sets by the Axiom of Regularity — i.e. if u is non-empty, then it contains a y such that $y \cap u = \emptyset$. A similar incompleteness result can be obtained via a more involved version of the argument illustrated by Proposition 4.4 below.

able to show that there is a term witnessing the existential quantifier in the statement $\exists y(U_1 \in y \wedge y \in U_1)$ appearing in the original proof. In light of this, we will work over $\mathsf{S}^+ = \mathsf{S} + V = L$ and consider an interpretation $(\cdot)^*$ of S^+ in $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+)$ obtained as in §4.2. Next consider the predicate

$$\beta_1(x) = x \in \omega \wedge \neg\exists y(\mathrm{f}(x) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(x))) \tag{4.10}$$

which is obtained from $\upsilon_2'(x)$ in the same manner that $\kappa(x)$ is obtained from $\rho'(x)$ in the proof of Proposition 4.3. As $\beta_1(x)$ is predicative, it follows that there is a class term B_1 such that $\mathsf{S} \vdash \forall x(x \in B_2 \leftrightarrow \beta_1(x))$ and $b_1 \in \omega$ such that $\mathsf{S} \vdash \mathrm{N}(b_1) = B_1$. We can now formulate the following:

PROPOSITION 4.4 *Suppose* $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+)$ *is consistent. Then* $\mathsf{S} \nvdash b_1 \in B_1$ *and* $\mathsf{S} \nvdash b_1 \notin B_1$.

Proof. As in the case of Proposition 4.3, $\mathsf{S}^+ \vdash B_1 \subseteq \omega$ and thus $\mathsf{S}^+ \vdash M(B_1)$. Applying definition (4.10), we obtain

$$\mathsf{S}^+ \vdash b_1 \in B_1 \leftrightarrow \neg\exists y(\mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(b_1))) \tag{4.11}$$

Now suppose that $\mathsf{S}^+ \vdash b_1 \in B_1$ and thus also $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+) \vdash (b_1 \in B_1)^*$. But then $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+) \vdash \mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(\mathrm{b}_1))$. But then trivially $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+) \vdash \mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(b_1)) \wedge \mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(b_1))$. Generalizing we thus get $\mathsf{S}^+ \vdash \exists y(\mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(b_1)))$ which is in turn equivalent to $\mathsf{S}^+ \vdash \neg(b_1 \in B_1)^*$. But this entails that $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+) \vdash \bot$.

On the other hand suppose that $\mathsf{S}^+ \vdash b_1 \notin B_1$. In this case, $\mathsf{S}^+ \vdash \exists y(\mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(b_1)))$. But now consider the formula $\psi(y) = \mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(b_1))$. As $\psi(x)$ is predicative, there is some class term T such that $\mathsf{S}^+ \vdash \forall x(\psi(x) \leftrightarrow x \in T) \wedge T \neq \emptyset$. But since such a class is nameable, it follows by Theorem 3.7 that $\mathsf{S}^+ \vdash \mathrm{As}(T) \in T$. From this it follows that

$$\mathsf{S}^+ \vdash \mathrm{f}(b_1) \in^* \mathrm{N}^*(\mathrm{f}(\mathrm{As}(T))) \wedge \mathrm{f}(\mathrm{As}(T)) \in^* \mathrm{N}^*(\mathrm{f}(b_1)) \tag{4.12}$$

From the second conjunct we obtain $\mathsf{S}^+ \vdash (\mathrm{As}(T) \in B_1)^*$ which is in turn equivalent to $\mathsf{S}^+ \vdash (\beta_1(\mathrm{As}(T)))^*$. But now note that by permuting the conjuncts of (4.12) and generalizing we also have

$$\mathsf{S}^+ \vdash \exists y(\mathrm{f}(\mathrm{As}(T)) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(\mathrm{As}(T)))) \tag{4.13}$$

But then also

$$\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}) \vdash (\exists y(\mathrm{f}(\mathrm{As}(T)) \in^* \mathrm{N}^*(\mathrm{f}(y)) \wedge \mathrm{f}(y) \in^* \mathrm{N}^*(\mathrm{f}(\mathrm{As}(T)))))^* \tag{4.14}$$

But this is equivalent to $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}) \vdash (\neg\beta_1(\mathrm{As}(T)))^*$. We thus again conclude that $\mathsf{S}^+ + \mathrm{Con}(\mathsf{S}^+) \vdash \bot$.[34] □

[34] Note that the need to work over $\mathsf{S} + V = L$ rather than S arises here because the reasoning embodied by steps iv)–vii) of the original paradox requires us to expand the definition of U_1 *twice* — the first time to conclude that there is some y such that $U_1 \in y \wedge y \in U_1$ and the second to conclude that y is such that $\exists z(z \in y \wedge y \in z)$. (This is also a feature of several other set theoretic paradoxes — inclusive of those of Burali-Forti and Mimiranoff — which appears to distinguish them from that of Russell.) When formalizing the sort of reasoning at issue in the arithmetized setting, we are thus faced with the need to refer to classes defined relative to two iterations of the interpretation $(\cdot)^*$. As the proof of Proposition 4.4 illustrates, one means of ensuring that we are able to keep track of classes across the models which are defined by $(\cdot)^*$ is to adopt an assumption such as $V = L$ which is sufficient to ensure that they are denoted by terms.

4.3. Semantic paradoxes As I will return to discuss in §5.2, the second version of Russell's paradox illustrates a moral about predicative definability which in turn can be understood to relate directly to the semantic paradoxes. It was, however, via a different route by which Kreisel (1953, p. 47) first hinted at the possibility of using a truth definition for arithmetic to obtain an incompleteness result in a manner similar to Proposition 4.3. Such a construction was first carried out by Wang (1955). But at this time the Arithmetized Completeness Theorem had only been stated in the form of Theorem 3.2 and not in the form of Theorem 3.4. For whereas the proof of the former result sketched above constructs the interpretation $(\cdot)^{\mathfrak{z}}$ from $\gamma(x)$ via (3.8), the original proof given by Hilbert & Bernays (1939, §4.2) from which Wang worked constructed the predicates defining the non-logical terms of $\mathcal{L}_{\mathsf{T}}$ one at a time in the manner of Gödel's original completeness proof.[35]

Wang's approach to the paradoxes may thus be understood as a direct continuation of the strategy initiated by Hilbert & Bernays in the *Grundlagen*. In particular, he first adapted the truth definition $\mathrm{Tr}(x)$ for $\mathcal{L}_{\mathsf{Z}}$ in $\mathcal{L}^2_{\mathsf{Z}}$ which Hilbert and Bernays had given in (1939, §5.2e) to a set theoretic system similar to S. He then employed Theorem 3.2 to obtain an interpretation $(\cdot)^*$ of S in $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ in the same manner as Kreisel (1950). In this way he obtained an interpretation $\mathrm{Tr}^*(x)$ of the truth predicate in the arithmetical fragment of $\mathcal{L}_{\mathsf{S}}$. Wang then went on to observe that when the primitive truth predicate $\mathfrak{W}(x)$ which Hilbert & Bernays had employed in their axiomatic treatment of the paradoxes in §5.1a is replaced with $\mathrm{Tr}^*(x)$, the Liar-like sentence constructed in the same manner becomes undecidable in S. This result can thus be understood to further bear out Hilbert & Bernays's prediction in (1939, §5.1c) that when the concepts involved in formulating the Liar are formalized mathematically, the paradox is transformed into a formal incompleteness result.

4.3.1. A second-order truth definition In order to reconstruct and generalize this method, it will be useful to work not over S but rather a theory of second-order arithmetic which is closer to that over which Hilbert & Bernays formulated their original truth definition. Many of the observations which are required to carry out this task pertain to the *expressibility* of the truth definition in $\mathcal{L}^2_{\mathsf{Z}}$ rather than the ability of a given theory to prove that the corresponding set of true $\mathcal{L}_{\mathsf{Z}}$-sentences exists. There is thus some flexibility in the choice of subsystems of Z_2 which may be used to carry out the formalization. However the theory ACA_0 suggests itself not only because it conservatively extends PA in the same manner as GB extends ZF, but also because it is similar to the system in which Hilbert & Bernays proposed to formalize analysis in Supplement IV of (1939).

Turning to the task of formulating the truth definition, recall first that already in $\mathcal{L}_{\mathsf{Z}}$ we can define an evaluation function $\mathrm{val}(x)$ for closed $\mathcal{L}_{\mathsf{Z}}$-terms such that $\mathsf{ACA}_0 \vdash \mathrm{val}(\ulcorner t \urcorner) = t$. Also let $\mathrm{ClTerm}_{\mathcal{L}_{\mathsf{Z}}}(x)$ formalize that x is the Gödel number of a closed term of $\mathcal{L}_{\mathsf{Z}}$ and Σ^0_y-$\mathrm{Sent}_{\mathcal{L}_{\mathsf{Z}}}(x)$ and Π^0_y-$\mathrm{Sent}_{\mathcal{L}_{\mathsf{Z}}}(x)$ respectively formalize that x is the Gödel number of a Σ^0_y- or Π^0_y-formula of $\mathcal{L}_{\mathsf{Z}}$ (where y is understood as a free variable). We may now construct an $\mathcal{L}^2_{\mathsf{Z}}$-formula $\tau(y, X)$ which expresses that the number class X is a truth definition for $\Sigma^0_y \cup \Pi^0_y$-sentences of $\mathcal{L}_{\mathsf{Z}}$ as follows:[36]

---

[35] Recall that if T and $\gamma(x)$ are as in §3.3, then $\gamma(x)$ can be understood as providing a truth definition for a model $\mathcal{M}_2$ which is strongly definable in $\mathcal{M}_1 \models \mathsf{T} + \mathrm{Con}(\mathsf{T})$ while on the other hand $(\cdot)^{\mathfrak{z}}$ only directly provides an interpretation of the atomic formulas of $\mathcal{L}_{\mathsf{T}}$. Thus while Theorem 3.4 can be understood as providing a characterization of the *elementary* diagram of $\mathcal{M}_1$, Theorem 3.2 only directly provides a characterization of its *atomic* diagram.

[36] Here expressions such as $x \cdot \ulcorner = \urcorner \cdot y$ are used to abbreviate the Gödel number of the sentence formed by concatenating the term Gödel numbered by x with the Gödel number of the equality sign and the term Gödel numbered by y. A similar truth definition is presented by Hilbert & Bernays (1939, p. 333-334/347). This may be

$$\begin{aligned}\tau(y,X) \quad =_{\mathrm{df}} \quad & \forall z((z \in X \to \Sigma^0_y\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(z) \vee \Pi^0_y\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(z)) \wedge \\ & \forall x\forall y(\mathrm{ClTerm}_{\mathcal{L}_\mathsf{Z}}(x) \wedge \mathrm{ClTerm}_{\mathcal{L}_\mathsf{Z}}(y) \to (x \cdot \ulcorner = \urcorner \cdot y \in X \leftrightarrow \mathrm{val}(x) = \mathrm{val}(y))) \wedge \\ & \forall x\forall y(\Sigma^0_{y-1}\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(x) \wedge \Sigma^0_y\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(y) \to (x \cdot \ulcorner \wedge \urcorner \cdot y \in X \leftrightarrow x \in X \wedge y \in X)) \wedge \\ & \forall x(\Sigma^0_y\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(x) \to (\ulcorner \neg \urcorner \cdot x \in X \leftrightarrow x \notin X)) \wedge \\ & \forall x\forall z(\mathrm{Var}(z) \wedge \Sigma^0_{y-1}\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(\ulcorner \forall \urcorner \cdot z \cdot x) \wedge \ulcorner \forall \urcorner \cdot z \cdot x \in X \leftrightarrow \forall v(\mathrm{subst}(x,v) \in X)) \wedge \\ & \forall x\forall z(\mathrm{Var}(z) \wedge \Pi^0_{y-1}\text{-Sent}_{\mathcal{L}_\mathsf{Z}}(\ulcorner \exists \urcorner \cdot z \cdot x) \wedge \ulcorner \exists \urcorner \cdot z \cdot x \in X \leftrightarrow \exists v(\mathrm{subst}(x,v) \in X)))\end{aligned}$$

We can now specify a Σ^1_1-formula of $\mathcal{L}^2_\mathsf{Z}$ which uniformly defines truth for $\mathcal{L}_\mathsf{Z}$-sentences as follows:

$$\mathrm{Tr}(x) =_{\mathrm{df}} \exists X \exists y(\tau(y,X) \wedge x \in X) \tag{4.15}$$

By an external induction on the complexity of formulas, it may now be shown that $\mathrm{Tr}(x)$ satisfies Tarski's Convention T in the sense that for all $\mathcal{L}_\mathsf{Z}$-sentences φ

$$\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner \varphi \urcorner) \leftrightarrow \varphi \tag{4.16}$$

This definition may additionally be shown to satisfy the standard Tarskian compositional clauses — i.e. for all $\mathcal{L}_\mathsf{Z}$-sentences φ, ψ

$$\mathrm{Tr}(\ulcorner \varphi \wedge \psi \urcorner) \leftrightarrow \mathrm{Tr}(\ulcorner \varphi \urcorner) \wedge \mathrm{Tr}(\ulcorner \psi \urcorner) \tag{4.17a}$$

$$\mathrm{Tr}(\ulcorner \neg \varphi \urcorner) \leftrightarrow \neg \mathrm{Tr}(\ulcorner \varphi \urcorner) \tag{4.17b}$$

$$\mathrm{Tr}(\ulcorner \forall x \varphi(x) \urcorner) \leftrightarrow \forall x \mathrm{Tr}(\mathrm{subst}(\ulcorner \varphi(v) \urcorner, x)) \tag{4.17c}$$

are provable in ACA_0.

These observations suggest both that the language $\mathcal{L}^2_\mathsf{Z}$ is sufficiently expressive to provide a natural formalization of first-order arithmetical truth and that the theory ACA_0 is sufficient to show that it demonstrates at least some of the adequacy conditions which are traditionally imposed on such definitions. I will return to discuss some additional desiderata which bear on such a determination in §5.1 and §5.5.

4.3.2. The Liar If $\gamma(x)$ is a formula defining a Henkin-complete extension of ACA_0 constructed in the manner of §3.3, then we have seen in §4.1 that it is possible to obtain an undecidable sentence λ_γ such that $\mathsf{Z} \vdash \lambda_\gamma \leftrightarrow \neg\gamma(\ulcorner \lambda_\gamma \urcorner)$ by applying the Diagonal Lemma directly to $\neg\gamma(x)$. We have also seen how $\gamma(x)$ provides a truth definition for a model $\mathcal{M}_2 \models \mathsf{ACA}_0$ which is strongly definable relative to another model $\mathcal{M}_1 \models \mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$. λ_γ may thus be likened to a Liar sentence in the sense that it is true in $\mathcal{M}_1$ if and only if it is false in $\mathcal{M}_2$.

The availability of the truth definition $\mathrm{Tr}(x)$ provides also another means of approaching the status of the Liar. For since the scope of (4.16) is limited to *first-order* sentences (i.e. those of $\mathcal{L}_\mathsf{Z}$), the existence of a *second-order* formula (i.e. one of $\mathcal{L}^2_\mathsf{Z}$) defining first-order arithmetical truth in the sense of (4.16) does not violate Tarski's undefinability

---

compared with the informal description of a formal truth definition for type theory previously provided by Tarski (1935) (see p. 195 and p. 255 in Tarski, 1956). Relative to the terminology which was adopted later, what Hilbert and Bernays showed is that *True Arithmetic* — i.e. the set TA of first-order statements true in the standard model of $\mathcal{L}_\mathsf{Z}$ – is *implicitly definable* in the sense that there exists an $\mathcal{L}^2_\mathsf{Z}$-formula with no bound second-order variables and a single free second-order variable X which is uniquely satisfied by TA in the standard model of $\mathcal{L}^2_\mathsf{Z}$. This can be shown to be equivalent to the fact that TA is Δ^1_1-definable in $\mathcal{L}^2_\mathsf{Z}$ — see, e.g., (Rogers, 1987, §15.1.XII). Kreisel & Wang (1955, p. 104) would also later cite Hilbert & Bernays's definition as a template for their subsequent construction of *partial* truth predicates $\mathrm{Tr}_{\Sigma_k}(x)$ for Σ^0_k-sentences in $\mathcal{L}_\mathsf{Z}$.

theorem. Nonetheless, it is still possible to construct a Liar-like sentence λ_{Tr} by applying the Diagonal Lemma to $\neg\mathrm{Tr}(x)$ so that

$$\mathsf{ACA}_0 \vdash \lambda_{\mathrm{Tr}} \leftrightarrow \neg\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}}\urcorner) \tag{4.18}$$

In order to assess the status of λ_{Tr}, something more must be said about what is meant by "the" Diagonal Lemma. In its simplest form, this result provides a means of constructing for every formula $\varphi(x)$ of the language of a theory T interpreting a weak fragment of arithmetic a sentence δ_φ such that $\mathsf{T} \vdash \delta_\varphi \leftrightarrow \varphi(\ulcorner\delta_\varphi\urcorner)$. There are a number of formally distinct means by which δ_φ can be effectively constructed from $\varphi(x)$.[37] But what matters for present purposes is how the syntactic complexity of δ_φ varies with that of $\varphi(x)$. In order to simplify the discussion, it will thus be useful to assume that this is achieved in the "canonical" manner by taking $\delta_\varphi = \varphi(\mathrm{subst}(\overline{m}, \overline{m}))$ where $m = \ulcorner\varphi(\mathrm{subst}(x, x))\urcorner$.[38]

In this case the sentence λ_{Tr} obtained by applying the diagonalization procedure to $\neg\mathrm{Tr}(x)$ will be equivalent to a Π^1_1-formula of $\mathcal{L}_\mathsf{Z}$.[39] And thus since λ_{Tr} can be taken to begin with the prefix "$\forall X$" it will not fall under $\mathrm{Form}^0_y(x)$ for any y. This purely syntactic observation can be formalized in ACA_0 from which it follows that $\mathsf{ACA}_0 \vdash \neg\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}}\urcorner)$ and thus also $\mathsf{ACA}_0 \vdash \lambda_{\mathrm{Tr}}$. But of course no inconsistency follows from these facts. For since λ_{Tr} is not an $\mathcal{L}_\mathsf{Z}$-sentence, the fact that $\mathsf{ACA}_0 \vdash \lambda_{\mathrm{Tr}}$ does not entail $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}}\urcorner)$ since (4.16) only holds for $\mathcal{L}_\mathsf{Z}$-sentences. And although we have obtained the sentence λ_{Tr} via a specific diagonalization procedure, it is also easy to see that no fixed point of $\neg\mathrm{Tr}(x)$ can be provably equivalent to an $\mathcal{L}_\mathsf{Z}$-sentence provided that ACA_0 is consistent.[40] As was originally observed by Kleene (1943, p. 64), in this way the Liar is mathematically repurposed as a means of demonstrating the non-collapse of the arithmetical and analytical hierarchies.

The situation is different when we first use Theorem 3.2 to obtain an arithmetical interpretation of $\mathcal{L}^2_\mathsf{Z}$ and then diagonalize on the image of the second-order truth predicate. To see this let $(\cdot)^* : \mathcal{L}^2_\mathsf{Z} \to \mathcal{L}_\mathsf{Z}$ be an interpretation of ACA_0 in $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ obtained in the manner of Theorem 3.2 and $\mathrm{f}(x)$ the definable embedding given by Proposition 3.5. Also recall that $\mathrm{Tr}^*(x)$ denotes the result of applying the inductive clauses in the definition of $(\cdot)^*$ to the formula $\mathrm{Tr}(x)$ while treating x as a free variable. By applying the Diagonal Lemma to $\mathrm{Tr}^*(\mathrm{f}(x))$ we can now obtain a sentence λ_{Tr^*} such that

$$\mathsf{ACA}_0 \vdash \lambda_{\mathrm{Tr}^*} \leftrightarrow \neg\mathrm{Tr}^*(\mathrm{f}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner)) \tag{4.19}$$

We can now record the following:

PROPOSITION 4.5 *If* $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ *is consistent, then* $\mathsf{ACA}_0 \nvdash \lambda_{\mathrm{Tr}^*}$ *and* $\mathsf{ACA}_0 \nvdash \neg\lambda_{\mathrm{Tr}^*}$.

Proof. Since the range of $(\cdot)^*$ is $\mathcal{L}_\mathsf{Z}$, it follows that $\neg\mathrm{Tr}^*(\mathrm{f}(x))$ is an arithmetical formula from which it follows that λ_{Tr^*} is an arithmetical sentence. We thus have that $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner) \leftrightarrow \lambda_{\mathrm{Tr}^*}$. Putting this together with (4.19) yields $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner) \leftrightarrow$

---

[37] See, e.g., (Halbach & Visser, 2014a) for a discussion of various alternatives.

[38] See, e.g., Smoryński, 1985, p. 6).

[39] This is immediate if we assume that the language includes the substitution function $\mathrm{subst}(x, y)$. But even without this assumption, λ_{Tr} can still be taken to be of the form $\exists z(\theta(\overline{k}, \overline{k}, y) \wedge \varphi(z))$ where $\theta(x, y, z)$ is a Σ^0_1-formula representing substitution and $k = \ulcorner\exists y(\theta(x, x, y) \wedge \varphi(y))\urcorner$. In this case λ_{Tr} will have the form $\exists z(\theta(\overline{k}, \overline{k}, z) \wedge \neg\mathrm{Tr}(z))$ which is provably equivalent to a Π^1_1-formula.

[40] For by an external induction on the lengths of proofs, we can show that if $\mathsf{ACA}_0 \vdash \chi_1 \leftrightarrow \chi_2$ then $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\chi_1 \leftrightarrow \chi_2\urcorner)$ and thus also $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\chi_1\urcorner) \leftrightarrow \mathrm{Tr}(\ulcorner\chi_2\urcorner)$. Thus if $\mathsf{ACA}_0 \vdash \psi \leftrightarrow \lambda_{\mathrm{Tr}}$ for $\psi \in \mathcal{L}_\mathsf{Z}$, then $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\psi\urcorner) \leftrightarrow \mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}}\urcorner)$. But now combining (4.18) and $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\psi\urcorner) \leftrightarrow \psi$ (since $\psi \in \mathcal{L}_\mathsf{Z}$), we would have $\mathsf{ACA}_0 \vdash \bot$.

$\neg\mathrm{Tr}^*(\mathrm{f}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner))$. But note that it also follows from (3.13) that $\mathsf{ACA}_0 \vdash \neg\mathrm{Tr}^*(\mathrm{f}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner)) \leftrightarrow (\neg\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner))^*$. We thus obtain $\mathsf{ACA}_0 \vdash \mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner) \leftrightarrow \neg(\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner))^*$. It thus follows by Corollary 4.2 that $\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}^*}\urcorner)$ — and hence also λ_{Tr^*} — are undecidable in ACA_0 under the assumption that $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ is consistent. □

4.3.3. The Grelling-Nelson Paradox The paradox concerning the notion *heterological* — i.e. the property possessed by an adjective which does not apply to itself — was first presented by Nelson & Grelling (1908). Ramsey (1926, p. 353) later grouped it along with the Liar amongst his list of "epistemological" paradoxes. On the other hand Russell (1903, p. 80) had already presented a similar contradiction by considering the notion of *predicability* — i.e. the property possessed by an expression just in case it is able to serve as a predicate. On this basis, both Nelson and Grelling (1908, pp. 60-61) and Hilbert (1917, pp. 193-194), (1928, pp. 890-891) presented this as leading to a paradox about sets which they explicitly liken to that of Russell.

Although I will develop this parallel further in §5.2, it may also be observed that such a classification is in the spirit of Hilbert & Bernays's treatment of the paradoxes in the *Grundlagen*. For as we have seen, they took the subject matter of what they termed "metamathematics" to subsume not only syntactic notions like provability, but also semantic ones such as truth and denotation. Although he does not mention the paradox by name, Kreisel (1950, p. 280-281) later observed in passing that reasoning similar to that of the Grelling-Nelson antinomy can be used to give an alternative reconstruction of Gödel's first incompleteness theorem.[41] But it should also not be surprising that this paradox can be directly converted into an incompleteness result by employing the method of arithmetization.

First note that once we have given a truth definition for $\mathcal{L}_\mathsf{Z}$ in $\mathcal{L}^2_\mathsf{Z}$, it is straightforward to also define the satisfaction relation for $\mathcal{L}_\mathsf{Z}$ formulas and closed terms. In particular, if $\varphi(x)$ is an $\mathcal{L}_\mathsf{Z}$-formula and t is a closed $\mathcal{L}_\mathsf{Z}$-term, then we can define $\mathrm{Sat}(\ulcorner\varphi(x)\urcorner, \ulcorner t\urcorner) =_{\mathrm{df}} \mathrm{Tr}(\mathrm{subst}(\ulcorner\varphi(x)\urcorner, \ulcorner t\urcorner))$. We then can show

$$\mathsf{ACA}_0 \vdash \mathrm{Sat}(\ulcorner\varphi(x)\urcorner, \ulcorner t\urcorner)) \leftrightarrow \varphi(t) \tag{4.20}$$

for all $\mathcal{L}_\mathsf{Z}$-formulas and closed terms t. This allows for a natural formalization of the predicate *heterological* as $\mathrm{Het}(x) =_{\mathrm{df}} \neg\mathrm{Sat}(x, x)$.

If we now let $h = \ulcorner\mathrm{Het}(x)\urcorner$ then the statement appearing in the Grelling-Nelson paradox is expressible as the $\mathcal{L}^2_\mathsf{Z}$-sentence $\mathrm{Het}(h)$ asserting that $\mathrm{Het}(x)$ does not hold of the numeral corresponding to its own Gödel number. But it should not come as a surprise that the paradox is averted in this case in the sense that $\mathsf{ACA}_0 \vdash \mathrm{Het}(\ulcorner h\urcorner)$. For note that this sentence is equivalent to $\neg\mathrm{Tr}(\ulcorner\neg\mathrm{Tr}(\ulcorner\mathrm{Het}(h)\urcorner)\urcorner)$. But this is in turn equivalent to the assertion that $\mathrm{Tr}(x)$ fails to hold of the Gödel number of a Π^1_1-formula – a fact which is again provable in ACA_0 simply in virtue of the syntactic form of $\neg\mathrm{Tr}(\ulcorner\mathrm{Het}(\ulcorner h\urcorner)\urcorner)$.

---

[41] For suppose we have set up a bijective Gödel numbering of $\mathcal{L}_\mathsf{Z}$-formulas with one free variable leading to an enumeration $\varphi_0(x), \varphi_1(x), \ldots$ Then the formula $\neg\mathrm{Prov}_\mathsf{Z}(\ulcorner\varphi_{\dot{x}}(\dot{x})\urcorner)$ will appear as $\varphi_q(x)$ for some $q \in \mathbb{N}$. It thus follows that $\varphi_q(\overline{q})$ is equivalent to $\varepsilon_\mathsf{Z} =_{\mathrm{df}} \neg\mathrm{Prov}_\mathsf{Z}(\ulcorner\neg\mathrm{Prov}_\mathsf{Z}(\ulcorner\varphi_{\dot{x}}(\dot{x})\urcorner)\urcorner)$ for which it can in turn be shown that $\mathsf{Z} \vdash \varepsilon_\mathsf{Z} \leftrightarrow \neg\mathrm{Prov}_\mathsf{Z}(\ulcorner\varepsilon_\mathsf{Z}\urcorner)$ – i.e. ε_Z is a *Gödel sentence* for Z and is thus undecidable in Z (presuming it is ω-consistent). When understood in this way, the Gödel sentence bears a natural resemblance to the statement arising in the Grelling-Nelson paradox. For say a formula $\varphi_n(x)$ is *autoprovable* if it is provable in Z that $\varphi_n(x)$ holds of $\ulcorner\varphi_n(x)\urcorner$ – i.e. $\mathrm{Prov}_\mathsf{Z}(\ulcorner\varphi_n(\ulcorner\varphi_n(x)\urcorner)\urcorner)$ – and *heteroprovable* otherwise. Then the statement ε_Z just described can be seen as asserting that the property of being heteroprovable is hetereprovable. But this is just another means of asserting that ε_Z expresses its own unprovability. See Kripke (2014) for a similar reconstruction.

But consider now the formula

$$\mathrm{Het}^*(x) = \neg\mathrm{Sat}^*(\mathrm{f}(x), \mathrm{f}(x)) \tag{4.21}$$

Letting $h^* = \ulcorner \mathrm{Het}^*(x) \urcorner$, we have the following:

PROPOSITION 4.6 *If* $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ *is consistent, then* $\mathsf{ACA}_0 \nvdash \mathrm{Het}^*(\ulcorner h^* \urcorner)$ *and* $\mathsf{ACA}_0 \nvdash \neg\mathrm{Het}^*(\ulcorner h^* \urcorner)$.

Proof. Note that $\mathrm{Het}^*(\ulcorner h^* \urcorner)$ is provably equivalent to $\neg\mathrm{Sat}^*(\mathrm{f}(\ulcorner \mathrm{Het}^*(x) \urcorner), \mathrm{f}(\ulcorner h^* \urcorner))$ by (4.21) and thus also to $(\neg\mathrm{Sat}(\ulcorner \mathrm{Het}^*(x) \urcorner, \ulcorner h^* \urcorner))^*$ by (3.13). On the other hand $\mathsf{ACA} \vdash \mathrm{Sat}(\ulcorner \mathrm{Het}^*(x) \urcorner, \ulcorner h^* \urcorner) \leftrightarrow \mathrm{Het}^*(\ulcorner h^* \urcorner)$ and thus $\mathsf{ACA}_0 \vdash (\neg\mathrm{Sat}(\ulcorner \mathrm{Het}^*(x) \urcorner, \ulcorner h^* \urcorner))^* \leftrightarrow \neg(\mathrm{Het}^*(\ulcorner h^* \urcorner))^*$. We thus obtain that $\mathsf{ACA}_0 \vdash \mathrm{Het}^*(\ulcorner h^* \urcorner) \leftrightarrow \neg(\mathrm{Het}^*(\ulcorner h^* \urcorner))^*$, from which it follows that $\mathrm{Het}^*(\ulcorner h^* \urcorner)$ is undecidable in ACA_0 if $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ is consistent by Corollary 4.2. □

§5. Discussion After presenting a result similar to Propopsition 4.3, Kreisel (1950, p. 280ff.) commented at length on its relationship to Gödel's original proof of the first incompleteness theorem. He begins as follows:

> Both the present paper and Gödel (1931b) use the diagonal (non-enumerability) argument to construct undecided propositions. Though this point is obvious, it seems worth mentioning: for one thing it connects undecidability proofs which are usually referred to as paradoxes and self references, with a familiar technique of mathematics, and, roughly speaking, allows one to convert non-enumerability proofs into those of undecidability. But also it throws light on the diagonal argument, and its "permissibility" e.g. Borel (1898).

In commenting on his generalizations of Kreisel's result, Wang (1955, p. 32) went a step further in relating formal incompleteness to different means of resolving the paradoxes:

> These results, it is hoped, will throw some new light on the semantic paradoxes. Moreover, like the result of Kreisel (1950), the conclusions are also of interest in that they connect indefinable classes and relations with undecidable sentences. There seems to be a certain similarity between this situation and the possibility of two different methods of avoiding mathematical paradoxes: the usual way of refusing to countenance a class R of all classes which are not self-members, and the alternative approach of admitting such a class R but treating as undecidable the question whether R belongs to R.

These passages suggest that both Kreisel and Wang regarded the paradoxes as originating from a common source which is not overtly metamathematical in character.[42] In

[42] Kreisel's allusion to Borel appears to refer to pp. 107-109 of (1898). In this section Borel first presents and then critiques Cantor's use of an enumeration of the characteristic functions of the subsets of a given set E in order to show its powerset is of cardinality greater than E. (Borel's complaint was that although Cantor's method can be used to show that the set of discontinuous functions on $\mathbb{R}$ is of cardinality greater than that of the continuum, it does not provide a means by which we can conceive of the elements of this set.) Kreisel additionally refers (1950, p. 280) to Hilbert & Bernays's use of diagonalization to show the existence of a total recursive function which is not primitive recursive (1934, p. 330/335) (and which is in fact already anticipated by Hilbert's (1926, pp. 388-391) discussion of the Ackermann's (1928) definition of a non-primitive recursive function in relation to the diagonal method). Kreisel additionally would also have most likely been aware of Lebesgue's (1905) and Lusin's (1925) use of

light of this, Wang's further proposal can be understood as offering a choice between two alternatives: i) on the one hand, it is possible to regard the paradoxes as arising from the misapplication of definitions which, when properly understood, fail to determine sets; ii) on the other hand, it is also possible to understand the definitions in question as determining classes but doing so in a manner such that paradoxical reasoning is blocked by the undecidability of particular statements about membership.

The results of §4 are indeed suggestive that formal undecidability might play a role in a uniform response to the paradoxes in the manner of Wang's second alternative. A thorough philosophical evaluation of this proposal is beyond the scope of the present paper. However in the remainder of this section, I will explore five questions which such an approach would need to confront were it to be further articulated: 1) How are the undecidability results in question related to Hilbert & Bernays's original concerns about proving the consistency of infinitary mathematics by finitary means? 2) Does the approach of §4 in fact lead to a unification of the set theoretic and semantic paradoxes? 3) Is it reasonable to regard the undecidable statements obtained there as genuinely indeterminate in truth value (which, as we will see, Kreisel appears to have thought)? 4) Should these results be regarded as *extensional* or *intensional* in the sense of Feferman (1960, §1)? 5) How do the methods and results developed in §4 interface with more recent work on axiomatic theories of truth?

5.1. Resolution? The developments surveyed in §2 – §4 illustrate a pathway from the paradoxes to the phenomenon of formal incompleteness. Hilbert & Bernays devoted much of the second volume of the *Grundlagen* to expositing Gödel's incompleteness theorems and investigating their interaction with various proof theoretic methods. In the first volume, however, it is the problem of ensuring that theories sufficient for the development of analysis are free from contradiction which sets the tone for their exposition of the finitary standpoint. In fact it was this concern which led them to propose the *method of arithmetization* described in §2:

> We are therefore forced to investigate the consistency of theoretical systems without considering actuality, and thus we find ourselves already at the standpoint of formal axiomatics. Now, one usually treats this problem — both in geometry and the disciplines of physics — with the *method of arithmetization*. The objects of a theory are represented by numbers or systems of numbers and the basic relations by equations or inequations, such that, on the basis of this translation, the axioms of the theory turn out either as arithmetical identities or provable sentences ... or as a system of conditions whose joint satisfiability can be demonstrated via arithmetical existence sentences. (1939, p. 3/3)

Comparison of Hilbert & Bernays's informal description of the method of arithmetization (1934, pp. 18-19/18-19) with their presentation of the Arithmetized Completeness Theorem (1939, pp. 234-253/243-263) suggests that they ultimately came to regard the latter as a mathematical embodiment of the former. And although no further use of this result is made in the *Grundlagen* itself, Bernays (1954a, §19) later presented a detailed construction of an arithmetical model of GB^- = GB - Infinty using techniques similar to those of §3.3. Although this theory does not entail the existence of infinite *sets*, it does prove the existence of infinite *classes* — e.g. that of all finite ordinals. And as is shown by Bernays (1942), it also allows for a development similar to that proposed at the end of the *Grundlagen*. As such, GB^- is typical of the theories which Hilbert & Bernays's spoke

---

diagonal arguments to demonstrate hierarchy theorems for point classes in descriptive set theory.

of as involving "idealization [which] ... transcends the realm of experience and intuitive self-evidence" (1939, p. 3/3).[43]

It is thus notable that Bernays described his construction of an arithmetical model of GB^- via the Henkin method as "establish[ing] in a constructive sense" (1954a, p. 93) the consistency of GB^- within "the frame of the number theoretic formal system Z of *Grundlagen der Mathematik* with certain additions" (1954a, p. 88) (where the "additions" in question consist of various extensions by definition). This terminology reflects on the one hand that the construction does indeed provide a relative consistency proof for GB^- in PA.[44] And on the other, it reflects that by this point Bernays had independently come to accept results in infinitary proof-theory as providing sufficient evidence to secure the consistency of first-order arithmetic.[45]

Bernays's comments on the method of arithmetization stand in contrast to the evaluation of the consistency proof for PA using the truth definition $\mathrm{Tr}(x)$ which is presented in the *Grundlagen* (1939, §5.2e). This is a rendition of the (now) familiar "semantic argument" for the consistency of PA – i.e. if it can be shown that the axioms of PA fall under $\mathrm{Tr}(x)$ and that the rules of first-order logic preserve truth, then it is also possible to demonstrate the global reflection principle for PA

$$\forall x((\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \wedge \mathrm{Prov}_{\mathsf{PA}}(x)) \rightarrow \mathrm{Tr}(x)) \tag{5.1}$$

and thus also its consistency.[46]

Hilbert & Bernays describe the semantic argument after their proof of the second incompleteness theorem in order to address the question of whether the introduction of "bound formula variables" is required to prove the consistency of PA (1939, p. 329/342). This argument can indeed be formalized in a system of second-order arithmetic in which it is possible to either prove the existence of a class corresponding to the extension of $\mathrm{Tr}(x)$ or in which induction can be directly performed on predicates defined in terms of this formula.[47] But as was observed by Mostowski (1950) in the set theoretic case, this

---

[43] This is how Hilbert & Bernays described the sorts of infinitary systems which arise in the axiomatization of geometry and physics alluded to in the passage cited above. The use of the expression "method of arithmetization" in the *Grundlagen* should thus not be understood simply as a redescription of Gödel's arithmetization of syntax (although it is reasonable to suppose Bernays came to view this as an *application* of the method). Rather it is intended to describe the more general use of arithmetical models which Hilbert had first employed in his consistency and independence proofs for various systems of geometry in (1899) and axiomatizations of physics in (1916). See also (Bernays, 1930, p. 253) for a related description.

[44] Recall that PA corresponds to the system which Hilbert & Bernays call Z. The full system is needed for Bernays's result whose proof is similar to the model expansion argument which was introduced by Novak (1950) to prove the relative consistency of GB over ZF.

[45] This topic is explored in (1939, §5.3) wherein Hilbert & Bernays present Gentzen's (1936) consistency proof in the context of considering potential extensions of the finitary standpoint as characterized at the beginning of the *Grundlagen*. It is also evident from Bernays's later philosophical writings (e.g. Bernays, 1954b) that he grew increasingly sympathetic to consistency proofs based on transfinite induction and other broadly constructive techniques.

[46] In fact Hilbert & Bernays (1939, p. 338/351-352) provide what is apparently the first explicit statement of the global reflection principle as well as sketching the semantic argument for PA in a system similar to second-order arithmetic. See (Dean, 2015) for discussion of the subsequent reception and significance of these developments.

[47] The theory Δ^1_1-CA_0 is sufficient relative to the former criterion. A theory which is sufficient relative to the latter is the system ACA in which the induction axiom of ACA_0 is replaced by the full-second order induction scheme.

argument cannot be used to prove the consistency of ZF in a second-order system such as GB which provides only predicative comprehension.[48] It is easy to see that the same is true for ACA_0 with respect to PA. Such an observation is in line with Hilbert & Bernays's remark that their proof of (5.1) is "not of finitary character" (1939, p. 338/351). And they went on to contend that the second-order definability of $\mathrm{Tr}(x)$ does not provide a means of determining whether this formula holds of arbitrary $\mathcal{L}_\mathsf{Z}$-sentences nor does it supply finitary justification for the application of *tertium non datur* to number theoretic reasoning (1939, pp. 338-339/351).

In regard to the latter point, Mostowski later observed that "since the 'whole theory of truth' makes it possible to prove the consistency of a system for which the notion of satisfaction has been defined, we infer that certain properties of the notion for [ZF] cannot be established in [GB]" (1950, p. 112). On the other hand, Mostowski also remarked that GB provides a "good" (p. 118) theory of truth in the sense that it allows us to derive all of the T-biconditionals (4.16) and compositional clauses (4.17) for $\mathrm{Tr}(x)$ as *schema* for $\mathcal{L}_\mathsf{Z}$-sentences. But the inability of such a theory to formalize the semantic argument is a manifestation of the fact that it does not prove the *uniform* version of the compositional clause for negation — i.e.

$$\forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \rightarrow (\mathrm{Tr}(\dot{\neg}x) \leftrightarrow \neg\mathrm{Tr}(x))) \tag{5.2}$$

As a consequence, ACA_0 also does not prove the uniform version of the law of the excluded middle with respect to $\mathrm{Tr}(x)$ — i.e. $\forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \rightarrow (\mathrm{Tr}(x) \vee \mathrm{Tr}(\dot{\neg}x)))$.

One might conclude on this basis that such systems provide "highly incomplete" theories of truth relative to the standards which Tarski (1956, p. 257) ultimately appears to have advocated.[49] But Hilbert & Bernays's investigation of the truth predicate was not motivated by a desire to study its (putative) role in informal reasoning in the style of more recent axiomatic theories of truth which I will discuss further in §5.5. For as we have just seen, they put little stock in the use of a second-order truth definition in a formal consistency proof. And on the other hand, we have also seen that their original presentation of the paradoxes in (1939, §5.1a) was not intended to highlight the role of the axiomatic approach as such but rather to serve the expository role of showing how the formalization of the Liar leads to incompleteness rather than formal inconsistency.

Hilbert & Bernays originally expanded on this point in (1939, §5.1b,c) by showing in the (now) familiar manner that the formalization of the proof of Gödel's first incompleteness theorem leads to his second incompleteness theorem. But we have seen that a similar moral can be reached more directly by applying the method of arithmetization to a theory like ACA_0 over which it is possible to formalize a truth definition for $\mathcal{L}_\mathsf{Z}$-sentences by a Σ^1_1-formula of $\mathcal{L}^2_\mathsf{Z}$. In particular, this leads to the consideration of the $\mathcal{L}_\mathsf{Z}$-formula $\mathrm{Tr}^*(x)$ obtained by applying an interpretation $(\cdot)^*$ of ACA_0 in (e.g.) $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0)$ which is in turn similar to that employed in Bernay's relative consistency proof for GB^-. As we have seen, the Liar-like sentence λ_{Tr^*} which may then be obtained by diagonalizing on $\mathrm{Tr}^*(x)$ is indeed undecidable in ACA_0 (and, *mutatis muntandis*, over stronger theories as well).

We thus reach the conclusion that when the method of arithmetization is applied to interpret the "idealizing assumption" embodied by the use of impredicative second-order quantification involved in the truth definition, the fear that an inconsistency in the style of the Liar will arise is abated by the undecidability of λ_{Tr^*}. Note that this point remains

[48] Tarski (1935, p. 317, p. 359) (p. 198, p. 237 in the English translation in (1956)) also mentions the possibility of providing a similar consistency proof using his own truth definition. But he later added a footnote (1956, p. 237) acknowledging Mostowski's result and the necessity of employing a metatheory which is sufficiently strong to formalize the relevant induction.

[49] See (Halbach, 2011, §3) for a recent re-endorsement of this view.

even if we take ourselves to lack finitary evidence for the consistency of a theory like ACA_0 or GB^-. For although we must assume the formal consistency statement for such a theory in order to apply the Arithmetized Completeness Theorem, doing so allows us to see how the antinomies which might originally have made us suspicious about consistency are transformed not into contradictions but rather instances of formal incompleteness.[50]

As we have seen, a similar route to Wang's second alternative may also be distilled from the set theoretic paradoxes by constructing arithmetical interpretations of systems similar to GB in the manner of Bernays's (1954a) proof. But at least two questions remain: 1) does such an approach point to a unification of the set theoretic and semantic paradoxes? 2) are the independent sentences which are constructed in this manner decidable on the basis of further mathematical considerations or do they represent a different or more abiding form of incompleteness?

5.2. Unification? The first step of the method outlined at the beginning of §4 for transforming paradoxes into incompleteness results was to show how the relevant "paradoxical notions" – e.g. that of non-self membership, ordinality, truth, or satisfaction — can be formalized in the language of a two-sorted theory such as GB or ACA_0. These systems make a distinction of *type* — e.g. between sets or numbers and classes — and *order* — e.g. between predicative and impredicative formulas. It is thus not surprising that the resolutions of Russell's paradox and the Liar which they provide are similar to the type-theoretic resolution originally envisioned by Russell (1908) and later extended by Tarski (1935) in terms of a hierarchy of metalanguages.[51]

To briefly review the initial parallels:

1) The predicate $\rho(x) = x \notin x$ defines a class R over GB. On the other hand, GB refutes the formula $M(R)$ asserting that R is a set. Thus although $\mathsf{GB} \vdash R \notin R$, this does not lead to a contradiction since the definition of R only allows us to infer $X \in R$ from $X \notin R$ in the case that X is a set.
2) The language of second-order arithmetic may be used as a metalanguage for that of first-order arithmetic in the sense that there is an $\mathcal{L}^2_{\mathsf{Z}}$-formula $\mathrm{Tr}(x)$ which provably satisfies Tarski's Convention T for $\mathcal{L}_{\mathsf{Z}}$-formulas over ACA_0. On the other hand, the Liar sentence λ_{Tr} for this predicate is not a $\mathcal{L}_{\mathsf{Z}}$ formula. Thus although $\mathsf{ACA}_0 \vdash \neg\mathrm{Tr}(\ulcorner\lambda_{\mathrm{Tr}}\urcorner)$, this does not lead to a contradiction since the definition of $\mathrm{Tr}(x)$ only allows us to infer $\neg\varphi$ from $\neg\mathrm{Tr}(\ulcorner\varphi\urcorner)$ in the case where $\varphi \in \mathcal{L}_{\mathsf{Z}}$.

The definability of the naming function $\mathrm{N}(x)$ for predicative classes discussed above allows us to extend these parallels in what is now likely to be an expected direction. Note first that we can modify our prior definition of $\mathrm{N}(x)$ to obtain a definition of a function $\mathrm{N}_a(x)$ which enumerates only *arithmetical classes* which may be proven to be subsets ω.

---

[50] Of course the question remains as to whether Hilbert or any of his interlocutors ever took seriously the possibility that the sort of theories of second order arithmetic in which they sought to formalize analysis might have been inconsistent. On one side of this issue was Weyl (1918, p. 21) who had famously remarked that impredicative definitions of real numbers trap us "in an endless circle, in absurdities and contradictions entirely analogous to Russell's well-known paradox ..." However neither in (Weyl, 1918) nor in (Weyl, 1919) does he describe a concrete means by which such an analogy might be exploited to yield a contradiction in a theory such as Z_2 or one of its subsystems. And in apparent response Hilbert (1922, vol 2., p. 1118) made clear just how dubious he regarded this claim: "[T]he paradoxes of set theory cannot be regarded as proving that the concept of a set of integers leads to contradictions. On the contrary: all our mathematical experience speaks for the correctness and consistency of this concept."

[51] See (Church, 1976) for further comparison of Russell's and Tarski's resolutions to the paradoxes.

In parallel, we can also set up an indexing of predicative $\mathcal{L}_{\mathsf{S}}$-formulas $\varphi_0(x), \varphi_1(x), \ldots$ defining arithmetical classes such that all $n \in \mathbb{N}$

$$\mathsf{S} \vdash \forall x(x \in \mathrm{N}_a(n) \leftrightarrow \varphi_n(x)) \tag{5.3}$$

Following Mostowski (1950) we can additionally carry out the definitions of truth and satisfaction given in §4.3 in S as well as in ACA_0. Putting these observations together allows us to define a satisfaction predicate $\mathrm{Sat}_{\mathsf{S}}(x, y)$ based on the indexing of formulas determined by $\mathrm{N}_a(x)$ such that

$$\mathsf{S} \vdash \mathrm{Sat}_{\mathsf{S}}(n, m) \leftrightarrow \varphi_n(m) \leftrightarrow m \in \mathrm{N}_a(n) \tag{5.4}$$

Now consider the predicate $\rho_a(x) = x \in \omega \wedge x \notin \mathrm{N}_a(x)$ and observe that $\mathsf{S} \vdash \forall x(\rho_a(x) \leftrightarrow x \in \omega \wedge \neg\mathrm{Sat}_{\mathsf{S}}(x, x))$. As was the case for the predicate $\rho'(x)$ considered in §4.2, if $\mathrm{N}_a(x)$ were predicatively definable, $\rho_a(x)$ would be as well. And from this it would also follow that $\rho_a(x) \equiv \varphi_j(x)$ for some $j \in \mathbb{N}$. In this case $\varphi_j(n)$ would hold just in case $\neg\mathrm{Sat}_{\mathsf{s}}(n, n)$ — i.e. if the predicate $\varphi_n(x)$ failed to hold of the number naming the class which it defined. But this is just another way of saying that $\varphi_j(x)$ is *heterological* in the sense defined in §4.3.3. By taking $n = j$ we would thus have $\mathsf{S} \vdash \mathrm{Sat}_{\mathsf{S}}(j, j) \leftrightarrow \neg\mathrm{Sat}_{\mathsf{S}}(j, j)$ by (5.4). This illustrates just how closely the Russell and Grelling-Nelson antinomies are related in the context of two-sorted set theories like S or GB which have the capacity to formalize the naming relation. But it also illustrates how the type-theoretic distinction between sets and classes imposed by these theories relates to the order-theoretic distinction between predicative and impredicative definitions.

In contrast to this, the Arithmetized Completeness Theorem can be viewed as collapsing the distinction between both order and type on which the many familiar resolutions of the paradoxes depend. By way of illustration, suppose that T is a theory of second-order arithmetic such as ACA_0 formulated over $\mathcal{L}^2_{\mathsf{Z}}$ and that $(\cdot)^* : \mathcal{L}^2_{\mathsf{T}} \to \mathcal{L}_{\mathsf{Z}}$ is an interpretation of T in $\mathsf{Z} + \mathrm{Con}(\mathsf{T})$ provided by Theorem 3.2. In this case, the image of an $\mathcal{L}^2_{\mathsf{Z}}$-formula φ — i.e. one which potentially contains bound second-order variables – will be a first-order $\mathcal{L}_{\mathsf{Z}}$-formula φ^*. In particular, quantifiers of the form $\forall X \ldots$ and $\exists X \ldots$ in φ will thus be replaced with first-order quantifiers restricted by an appropriate domain predicate of the form $\forall x(\delta_2(x) \to \ldots)$ and $\exists x(\delta_2(x) \wedge \ldots)$ in φ^*.[52]

Recall that in three of the four cases we have considered, the statements shown to be undecidable via arithmetized completeness constructions are provably equivalent over T to the negation of their own interpretation under $(\cdot)^*$ — i.e. they are instances of the schema

$$\mathsf{T} \vdash \varphi \leftrightarrow \neg\varphi^* \tag{5.5}$$

[52] ACA_0 can be viewed as representing the first in a hierarchy of systems defined by adjoining to PA comprehension principles for formulas with bound variables ranging only over numbers (which determine *level one classes*), with bound variables ranging only over level one classes (which determine *level two classes*), ... the union of which Church (1956, §58) refers to as *ramified second-order arithmetic*. In this setting it is possible to draw a partial analogy between the function served by Russell's Axiom of Reducibility and the statement of the Arithmetized Completeness Theorem. The relevant version of the former would state that for every impredicatively defined level-n class, there exists a coextensive level 1-class. On the other hand, if $(\cdot)^* : \mathcal{L}^2_{\mathsf{T}} \to \mathcal{L}_{\mathsf{Z}}$ is defined as above, then the interpretation of an impredicative $\mathcal{L}^2_{\mathsf{Z}}$ formula $\varphi(x)$ defining a set $A \subseteq \mathbb{N}$, there will be an $\mathcal{L}_{\mathsf{Z}}$-formula $\varphi^*(x)$ defining a set $A^* \subseteq \mathbb{N}$. It is easy to see that A and A^* must coincide extensionally for all $n \in \mathbb{N}$ for which ACA_0 *decides* the truth value of $\varphi(\overline{n})$. But it follows from the existence of sets K from Proposition 4.3 that there will be cases in which A and A^* are provably *non-coextensive*.

This can be understood to reflect the conventional classification of the Russell, Liar, and Grelling-Nelson antinomies as paradoxes of self-reference.[53] But via arithmetization it is also possible to understand these paradoxes semantically in the sense that $(\cdot)^*$ strongly defines a model of $\mathcal{M}_2 \models \mathsf{T}$ relative to a model $\mathcal{M}_1 \models \mathsf{T} + \mathrm{Con}(\mathsf{T})$. As we have seen, the relevant undecidable statements can also be interpreted as asserting of themselves that they are true in $\mathcal{M}_1$ if and only if they are false in $\mathcal{M}_2$. And thus if we are willing to assume $\mathrm{Con}(\mathsf{T})$, the Arithmetized Completeness Theorem can also be understood as showing that it is possible to describe consistent situations in which the statements on either side of the biconditionals are true.

This also illustrates the sort of dissolution of the paradoxes in virtue of undecidability which Wang appears to have envisioned. Suppose, for instance, that the Liar sentence is understood to arise not from a defined formula like $\mathrm{Tr}(x)$, but rather from a primitive predicate $\mathtt{T}(x)$ intended to express the "everyday" or "naive" notion of truth about which we are attempting to reason axiomatically. As I will discuss further in §5.5, it is often said in this context that there is no reason to prefer the truth of one side of the resulting Liar biconditional $\lambda_{\mathtt{T}}$ over the other and thus also no reason to expect that a theory of truth should decide the truth value of $\lambda_{\mathtt{T}}$. This suggests that the undecidability of λ_{Tr^*} is consistent with our expectations about the resolution of the Liar.[54]

In the case of Russell's paradox, however, the iterative conception of set partially codified by the Axiom of Regularity gives us independent grounds for rejecting the existence of self-membered sets and thus for deciding in favor of $R \notin R$ over $R \in R$ (the non-sethood of R notwithstanding). But as we have seen, Kreisel's original incompleteness result concerns not R, but rather an arithmetized analog of the collection R' consisting of those numbers which are not members of the classes which they name. We have seen that no predicative definition of the collection R' is possible. But if we additionally attempt to probe our "naive" intuitions about the nameability relation, it would seem that there is no abiding reason to think that the number which names R' (were one to exist) either should or should not be a member of R'. And as I will now discuss, this appears to conform with what Kreisel himself thought about the status of the undecidable sentence $k \in K$ obtained in Proposition 4.3.

5.3. Absolute undecidability? Kreisel (1950, p. 267) originally formulated a version of Proposition 4.3 in the course of attempting to confirm a conjecture of Bernays (Hilbert & Bernays, 1939, p. 191/199) — i.e. that the predicate calculus is not complete with respect to models in which all predicates are computable [*berechenbare*], or as we would now say, Δ^0_1-definable. This informed his discussion of the distinction between the Gödel sentence δ_{Z} for an arithmetical theory Z such as PA and the undecidable Kreisel sentence $k \in K$ obtained in Proposition 4.3 over a system similar to S. We would now

---

[53] On the other hand, the current setting highlights the additional complications required to formalize the reasoning of antinomies such as the paradoxes of reciprocated or grounded classes which are themselves simplifications of the Burali-Forti and Mirimanoff paradoxes. *Pace* Priest (1994), this suggests that the classification of these antinomies as paradoxes of self-reference is in need of further refinement.

[54] Wang (1963, p. 386) would later liken his "second approach" to the paradoxes to the proposal of Fitch (1952) of blocking the reasoning of Russell's paradox by denying that the law of the excluded middle holds for the statement $R \in R$ (see §18). Fitch's proposal is based on his claim that both $R \in R$ and the Liar express "indefinite propositions" which lack truth values. But his account of why this is so (1952, §1.4) amounts to little more than the observation that were these statements to possess a definite truth value then a contradiction would be derivable in the calculus he describes. While this approach is also reminiscent of that adopted by certain axiomatic theories of truth (see §5.5), Fitch's proposal can also be contrasted with the one proposed here in virtue of failing to provide an independent explanation of the relevant failure of excluded middle.

express his point by saying that while δ_{Z} is a Π^0_1-sentence, $k \in K$ can be taken to have the form of either a Σ^0_2 or a Π^0_2-sentence. This follows immediately from the biconditional (4.6) together with the fact that Theorem 3.2 yields a Δ^0_2 definition of $\in^*$. But whichever form we take $k \in K$ to have, it cannot be equivalent to a Δ^0_1-formula as all such statements are decidable in Z.[55]

On the other hand, recall that δ_{Z} is provably equivalent over Z to $\forall x \neg \mathrm{Proof}_{\mathsf{Z}}(x, \ulcorner \delta_{\mathsf{Z}} \urcorner)$ which is a Π^0_1 statement. Gödel's proof of the first incompleteness theorem shows that if Z is ω-consistent then $\mathsf{Z} \nvdash \neg\delta_{\mathsf{Z}}$. But as Kreisel (1950, p. 278) observed, δ_{Z} must then be *true* in the standard model $\mathcal{N}$.[56] However Kreisel also observed that there is no evident way to extend this familiar observation beyond the class of Π^0_1-statements to argue from the undecidability of a Π^0_2- or Σ^0_2-sentence like $k \in K$ to its truth or falsity. On this basis he suggested that predicates defining classes such as K should be understood as "not [making] a division of integers into two classes T, F but rather three T, F, U" (1950, p. 283). Kreisel echoed more explicitly the view that *class* (as opposed to *set*) membership need not be understood as bivalent in (1967, p. 143) in the context of formulating a theory of sets and what he refers to as *definable properties* which is otherwise similar to S (pp. 162-165). And he later went so far as to remark that "nobody has any idea whether $[k \in K]$ is true or false" (Kreisel, 1969, p. 112).[57]

In order to understand what might be at stake with such pronouncements, it is useful to recall that Kreisel (1967, pp. 151-152) himself proposed a *three-way* classification of formally undecidable statements. The first class contains metamathematical Π^0_1-statements like δ_{Z} whose undecidability entails their truth via the sort of argument just rehearsed. The second class contains the Continuum Hypothesis which Kreisel (1967, p. 150) famously argued is such that either it or its negation is a second-order logical truth. And the third contains examples like the Parallel Postulate which he observed are such that their truth values are not fixed by adopting a second-order formulation of the geometric axioms.[58]

To take seriously Kreisel's apparent proposal about the non-bivalence of class membership or Wang's related suggestion that the paradoxes can be avoided by "admitting paradoxical classes" but "treating as undecided" questions about membership within them, more still needs to be said about the formal properties of the undecidable sentences constructed in §4. As I will suggest below, such statements do not fit comfortably into any of Kreisel's categories. In order to see why this is so it will be useful to first highlight several respects in which they are similar to prototypical members of the first class. But I will then suggest that they also have an important affinity with both the second and third.

5.4. Intensionality Feferman (1960) originally classified Gödel's first incompleteness theorem for PA as an *extensional* result in the sense that the undecidability of its Gödel

---

[55] Kreisel (1950, p. 275) claimed on this basis that Proposition 4.3 is already sufficient to provide an answer to Hilbert & Bernays's question. But as Wang (1953) later observed, the fact that the specific predicate $\in^*$ constructed in this proof yields an undecidable statement is not on its own sufficient to show that there does not exist an arithmetical model of S in which the interpretation of $\in$ is recursive. The non-existence of such an interpretation was later shown by Rabin (1958) using an argument involving recursively inseparable sets similar to the one which is now typically employed in the proof of Tennenbaum's Theorem (e.g. Kaye, 1991, §11).

[56] For if it were false, then $\mathcal{N} \models \mathrm{Proof}_{\mathsf{Z}}(\overline{n}, \ulcorner \delta_{\mathsf{Z}} \urcorner)$ for some $n \in \mathbb{N}$. But since Z is sufficiently strong to prove all true Δ^0_0-statements, then $\mathsf{Z} \vdash \mathrm{Proof}_{\mathsf{Z}}(\overline{n}, \ulcorner \delta_{\mathsf{Z}} \urcorner)$ and thus $\mathsf{Z} \vdash \neg\delta_{\mathsf{Z}}$.

[57] See, (Manevitz & Stavi, 1980, p. 146) and (Isaacson, 2011, p. 54) for similar assessments which anticipate the argument of §5.4.

[58] I.e. when the Continuity Axioms — V in (Hilbert, 1899) — are formalized as second-order statements rather than first-order schemes.

sentence δ_{PA} only depends on the fact that the arithmetical formula employed to formalize provability weakly represents derivability from its axioms. But he also famously illustrated a sense in which Gödel's second theorem is *intensional* by showing that in addition to the "canonical" consistency statement $\mathrm{Con}(\mathsf{PA})$ — for which $\mathsf{PA} \nvdash \mathrm{Con}(\mathsf{PA})$ as long as is PA consistent — it is also possible to construct a formula $\mathrm{Con}'(\mathsf{PA})$ expressing consistency relative to a different binumeration of the axioms of PA such that $\mathsf{PA} \vdash \mathrm{Con}'(\mathsf{PA})$.

Other parameters which have the potential for introducing similar sensitivities into the formulation of undecidability results include the following:

i) the arithmetical encoding which is chosen to Gödel number terms and formulas;
ii) the chosen form of the Diagonal Lemma which is employed to generate "self-referential" statements;
iii) the definitions of proof and provability on which the definition of formulas such as $\mathrm{Con}(\mathsf{T})$ are based.[59]

It is thus also natural to ask how the intensional aspects of arithmetization bear on the results of §4.

For concreteness, let T be a recursively axiomatizable $\mathcal{L}_{\mathsf{Z}}$-theory extending $\mathrm{I}\Sigma_2$. The simplest example of an undecidable statement we have considered is the Liar-like sentence λ_γ constructed in Lemma 4.1 by diagonalizing directly on the formula $\gamma(x)$ which defines a Henkin-completion of T. By examining the details of this construction, it is easy to see that parameters i) – iii) all contribute to the identity of λ_γ and thus also to that of the other undecidable statements constructed in §4. But this is also true of a number of parameters specific to the formalization of the Henkin construction on which the definition of $\gamma(x)$ is based. A non-exhaustive list of these is as follows:

iv) the extension of the Gödel numbering $\ulcorner \cdot \urcorner$ to the Henkin language $\mathcal{L}_{\mathsf{T}}^H$;
v) the enumeration $\psi_0, \psi_1, \ldots$ of $\mathcal{L}_{\mathsf{T}}^H$-sentences which are used on the inductive construction of the sets Γ_{T}^i;
vi) the method by which representatives from equivalence classes of Henkin constants up to provable equality in Γ_{T} are selected;
vii) the extension condition by which the set $\Gamma_{\mathsf{T}}^{i+1}$ is defined relative to Γ_{T}^i;
viii) the manner in which the formula $\gamma(x)$ formalizing membership in Γ_{T} is itself defined on the basis of the other parameters.

In order to gauge the extent to which these parameters bear on the results of §4, consider an $\mathcal{L}_{\mathsf{T}}$-formula $\rho(x)$ satisfying the following derivability-like conditions:

(5.6a) If $\mathsf{T} \vdash \varphi$, then $\mathsf{T} + \mathrm{Con}(\mathsf{T}) \vdash \rho(\ulcorner \varphi \urcorner)$.

(5.6b) $\mathsf{T} \vdash \rho(\ulcorner \neg \varphi \urcorner) \leftrightarrow \neg \rho(\ulcorner \varphi \urcorner)$

By mimicking the reasoning of Lemma 4.1 we obtain

PROPOSITION 5.1 *Suppose that $\rho(x)$ satisfies* (5.6a,b) *and that β is such that $\mathsf{T} \vdash \beta \leftrightarrow \neg\rho(\ulcorner \beta \urcorner)$. If $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ is consistent, then β is undecidable in T.*

It is not hard to see that if $\gamma(x)$ is defined as in §3.3 it will satisfy (5.6a,b) as long as the formulas $\mathrm{Prov}_{\mathsf{T}}(x)$ and $\mathrm{Con}(\mathsf{T})$ are defined canonically — in fact these properties are consequences of (3.7a,e). And from this it follows that the undecidability of the formula

[59] See (Halbach & Visser, 2014a,b) for a systematic study of how varying any of i) - iii) can affect the provability or refutability of Henkin or truth-teller sentences for partial truth predicates.

λ_γ obtained in Corollary 4.2 will not depend on any of the parameters i)–ii) and iv)–x).[60] In virtue of (4.1c), this observation extends to sentences such as $k \in K$, λ_{Tr^*}, or $\mathrm{Het}^*(h^*)$ which are provably equivalent to their negations under the interpretation $(\cdot)^*$. Thus although the method by which these statements are shown to be undecidable differs from that by which Gödel originally showed the undecidability of δ_{Z}, the results of §4 are also reasonably invariant with respect to the intensional parameters enumerated above.

On the other hand, it is also easy to see that statements like λ_γ do not resemble Gödel sentences in other respects. For suppose that T, $\mathrm{Prov}_{\mathsf{T}}(x)$, and $\mathrm{Con}(\mathsf{T})$ are defined canonically. Another familiar observation is that if φ is a fixed point of the formula $\neg\mathrm{Prov}_{\mathsf{T}}(x)$ — i.e. $\mathsf{T} \vdash \varphi \leftrightarrow \neg\mathrm{Prov}_{\mathsf{T}}(\ulcorner\varphi\urcorner)$ — then $\mathsf{T} \vdash \varphi \leftrightarrow \mathrm{Con}(\mathsf{T})$. Thus all Gödel sentences for T are provably equivalent to the canonical consistency statement for T and hence to one another.

But now consider the predicate $\neg\gamma(x)$. As we have seen, the Diagonal Lemma yields the existence of a sentence $\lambda_{\neg\gamma}$ such that $\mathsf{T} \vdash \lambda_{\neg\gamma} \leftrightarrow \neg\gamma(\ulcorner\lambda_{\neg\gamma}\urcorner)$. While both $\gamma(x)$ and $\mathrm{Prov}_{\mathsf{T}}(x)$ satisfy (5.6a), only $\gamma(x)$ satisfies (5.6b). But in virtue of this we have $\mathsf{T} \vdash \neg\gamma(\ulcorner\lambda_{\neg\gamma}\urcorner) \leftrightarrow \gamma(\ulcorner\neg\lambda_{\neg\gamma}\urcorner)$ and thus also $\mathsf{T} \vdash \neg\lambda_{\neg\gamma} \leftrightarrow \neg\gamma(\ulcorner\neg\lambda_{\neg\gamma}\urcorner)$. It hence follows that $\neg\lambda_{\neg\gamma}$ is another provable fixed point of $\neg\gamma(x)$ which is not provably equivalent to $\lambda_{\neg\gamma}$ (unless T is inconsistent).

These considerations suggest that the undecidable statements constructed in §4 differ from traditional Gödel sentences in that they lack what can reasonably be described as a *canonical form*. This leaves open the possibility that their truth values in the standard model are also not fixed independently of the decisions we make about the various intensional parameters. That this is indeed the case can be demonstrated by investigating further the role of parameter v) — i.e. the selection of an enumeration $\psi_0, \psi_1, \ldots$ of $\mathcal{L}^H_{\mathsf{T}}$-sentences — in the formalization of the Henkin procedure.

Suppose that we have fixed choices for parameters i) and iv) so that we have a Gödel numbering of $\mathcal{L}^H_{\mathsf{T}}$ and that the enumeration in question is also given by some fixed recursive function $\phi_j(x)$ such that $\phi_j(0) = \ulcorner\psi_0\urcorner, \phi_j(1) = \ulcorner\psi_1\urcorner, \ldots$ Additionally suppose that we have fixed parameter vii) in accordance with (3.6) such that a formula ψ_i is added to the set Γ^i_{T} just in case $\Gamma^i_{\mathsf{T}} \cup \{\psi_i\}$ is consistent (as in Henkin, 1949). Now suppose that φ is undecidable in T and that φ is also the first sentence in the enumeration — i.e. $\varphi = \psi_0$. Since in this case $\mathsf{T} \nvdash \neg\varphi$, we will thus have that $\varphi \in \Gamma^1_{\mathsf{T}}$ and thus also $\varphi \in \Gamma_{\mathsf{T}}$. If $\gamma(x)$ formalizes the definition of Γ_{T} based on these choices, then we will also have $\mathcal{N} \models \gamma(\ulcorner\varphi\urcorner)$ — i.e. when our formalization of the Henkin construction is carried out relative to the standard model, φ will be included in the maximally consistent set of sentences defined by $\gamma(x)$. By parallel reasoning, if it turned out that $\neg\varphi = \psi_0$, then we would have that $\mathcal{N} \models \gamma(\ulcorner\neg\varphi\urcorner)$ and thus also $\mathcal{N} \models \neg\gamma(\ulcorner\varphi\urcorner)$.[61]

---

[60] This is in conformity with Feferman's (1960, p. 39) original characterization of Theorem 3.2 as a result of "mixed extensional and intensional type". For as he observes, while the proof does depend on the fact that $\mathrm{Con}(\mathsf{T})$ is defined canonically, it does not depend on the formula we have chosen to binumerate the axioms of T in defining the predicate $\mathrm{Prov}_{\mathsf{T}}(x)$. Note, however, that this is a parameter which we do not otherwise need to be concerned with in the case where T is a finitely axiomatizable theory such as GB or ACA_0.

[61] Manevitz & Stavi (1980) similarly observed that by defining $\mathrm{Con}^1(\mathsf{T}) = \mathrm{Con}(\mathsf{T})$ and $\mathrm{Con}^{n+1}(\mathsf{T}) = \mathrm{Con}(\mathsf{T} + \mathrm{Con}^n(\mathsf{T}))$ and defining an enumeration which begins with $\mathrm{Con}^1(\mathsf{T}), \mathrm{Con}^2(\mathsf{T}), \mathrm{Con}^{n-1}(\mathsf{T}), \neg\mathrm{Con}^n(\mathsf{T})$ we are able to control the number of times which certain definitions of $\gamma(x)$ may be iterated before the set defined by the iterates $\gamma^0(x) = \gamma(x), \gamma^{i+1}(x) = \gamma^i(\ulcorner\gamma(\dot{x})\urcorner)$ becomes inconsistent. On the other hand, they also showed (1980, §3) that there exist other definitions of $\gamma(x)$ (which can even be Δ^0_2) whose iterates define consistent sets for all $i \in \mathbb{N}$. As Kreisel's "model theoretic" proof of the second incompleteness theorem (as presented by Smorynski, 1977, §6) depends

Next let $f_0(x,y)$ and $f_1(x,y)$ be recursive functions such that if $\phi_j(x)$ determines an enumeration of sentences as above, then $\phi_{f_0(j,k)}(x)$ determines the enumeration which lists ψ_k first and then continues as $\phi_j(x)$ but omitting ψ_k — i.e. $\psi_k, \psi_0, \psi_1, \ldots, \psi_{k-1}, \psi_{k+1}, \ldots$ (presuming $k > 0$) – and $\phi_{f_1(j,k)}(x)$ determines the enumeration which lists $\neg\psi_k$ first and then continues as $\phi_j(x)$ — i.e. $\neg\psi_k, \psi_0, \psi_1, \ldots, \psi_{k-1}, \psi_{k+1}, \ldots$ In cases where the undecidability of a given sentence φ does not depend on the choice of an enumeration of $\mathcal{L}^H_{\mathsf{T}}$ itself (e.g. Gödel sentences), these definitions allow us to uniformly define Henkin-completions of T in which $\gamma(\ulcorner\varphi\urcorner)$ is either true or false in $\mathcal{N}$ simply by fixing the stage at which φ is enumerated.

On the other hand, the construction of λ_γ clearly does depend on the definition of $\gamma(x)$ and hence on the chosen enumeration. Nonetheless, we can obtain the following:

PROPOSITION 5.2 *Let T be as above. Then there exist $\mathcal{L}_{\mathsf{T}}$-formulas $\gamma_0(x)$ and $\gamma_1(x)$ defining Henkin-complete extensions of T such that the statements λ_{γ_0} and λ_{γ_1} respectively obtained by applying the Diagonal Lemma to $\neg\gamma_0(x)$ and $\neg\gamma_1(x)$ have the following properties*: i) *if $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ is consistent, then λ_{γ_0} and λ_{γ_1} are both undecidable in T*; ii) *λ_{γ_0} is false in $\mathcal{N}$*; iii) *λ_{γ_1} is true in $\mathcal{N}$.*

Proof. We have just seen that it is possible to parameterize the definition of the formula $\gamma_i(x)$ defining a Henkin-complete extension of T based on the extension condition (3.6) in the index i of a given enumeration — say $\phi_i(x) = \psi_0, \psi_1, \ldots$ — of $\mathcal{L}^H_{\mathsf{T}}$-formulas. Since a standard formalization of the Diagonal Lemma allows us to effectively construct a fixed-point formula for $\gamma_i(x)$, there is also a recursive function $g(i)$ such that $\mathsf{T} \vdash \psi_{g(i)} \leftrightarrow \neg\gamma_i(\ulcorner\psi_{g(i)}\urcorner)$. We may hence consider the enumerations $\phi_{f_0(i,g(i))}(x)$ and $\phi_{f_1(i,g(i))}(x)$ whose first elements are respectively $\psi_{g(i)}$ and $\neg\psi_{g(i)}$. By the Recursion Theorem (e.g. Rogers, 1987, §11.2) there exist $n_0, n_1 \in \mathbb{N}$ such that $\phi_{f_0(n_0,g(n_0))}(x) = \phi_{n_0}(x)$ and $\phi_{f_1(n_1,g(n_1))}(x) = \phi_{n_1}(x)$. $\phi_{n_0}(x)$ and $\phi_{n_1}(x)$ hence respectively define enumerations beginning with $\psi_{g(n_0)}$ and $\neg\psi_{g(n_1)}$ and which are also such that

$$\mathsf{T} \vdash \psi_{g(n_0)} \leftrightarrow \neg\gamma_{n_0}(\ulcorner\psi_{g(n_0)}\urcorner) \tag{5.7a}$$

$$\mathsf{T} \vdash \psi_{g(n_1)} \leftrightarrow \neg\gamma_{n_1}(\ulcorner\psi_{g(n_1)}\urcorner) \tag{5.7b}$$

Since $\gamma_{n_0}(x)$ and $\gamma_{n_1}(x)$ satisfy (5.6a,b), $\lambda_0 =_{\mathrm{df}} \psi_{g(n_0)}$ and $\lambda_1 =_{\mathrm{df}} \psi_{g(n_1)}$ will both be undecidable in T provided that $\mathsf{T} + \mathrm{Con}(\mathsf{T})$ is consistent. Since λ_0 and $\neg\lambda_1$ occur first in the enumerations respectively defined by $\phi_{n_0}(x)$ and $\phi_{n_1}(x)$, we will hence have that $\mathcal{N} \models \gamma_{n_0}(\ulcorner\lambda_0\urcorner)$ and $\mathcal{N} \models \gamma_{n_1}(\ulcorner\neg\lambda_1\urcorner)$ and thus also $\mathcal{N} \models \neg\gamma_{n_1}(\ulcorner\lambda_1\urcorner)$. But now by (5.7a,b), we also have $\mathcal{N} \models \neg\lambda_0$ and $\mathcal{N} \models \lambda_1$. □

Proposition 5.2 shows that the truth value of λ_γ is indeed not fixed independently of the definition of the Henkin-completion $\gamma(x)$ relative to which it is defined. But one might still attempt to fix the truth value of such a statement by restricting attention to definitions of $\gamma(x)$ satisfying conditions which might be argued to characterize "canonical" formalizations of the Henkin procedure. In fact Kreisel (1965, §3.222) provided a precedent for this approach by proposing that we should regard an arithmetical formula $\mathrm{Prov}_{\mathsf{T}}$ as "canonically representing" provability in a theory T if its structure reflects that of the inductive definition of the derivability relation $\mathsf{T} \vdash \varphi$.[62]

Pace Kreisel, Halbach & Visser (2014a) provide a number of examples which illustrate that such judgements of "resemblance" between metatheoretic and object language definitions are difficult to make precise, even in the case of provability. As we have seen, a

---

on the fact that $\gamma^i(x)$ defines an inconsistent set for some i, this illustrates a sense in which this result is also intensional in Feferman's sense.

[62] Such a definition must evidently be Σ^0_1. Kreisel's criterion is thus sufficient to exclude from consideration the provability predicate involved in Feferman's (1960) definition of $\mathrm{Con}'(\mathsf{PA})$ mentioned at the beginning of this section which is Δ^0_2.

formalization of the Henkin procedure must build on such a definition while also contending with parameters iv) – viii). Several related considerations would seem to diminish even further the prospects of defining a notion of canonicity for arithmetical definitions of completions of T in a manner which would allow us to non-stipulatively determine the truth values of sentences like λ_γ:

1) Let $\Gamma = \{\varphi \mid \mathcal{N} \models \gamma(\ulcorner\varphi\urcorner)\}$ be the set of sentences defined by $\gamma(x)$ relative to the standard model. By definition, $\Gamma \supseteq \mathsf{T}$ and is consistent (as long as T is). But since $\gamma(x)$ is an arithmetical sentence, we cannot have $\Gamma = TA$ (i.e. True Arithmetic) in light of Tarski's Theorem. Hence any candidate for $\gamma(x)$ must incorrectly decides some arithmetical sentence.
2) The sequence of definitions given in §3.3 gives rise to a definition of $\gamma(x)$ which is Δ^0_2. But in fact this establishes neither an upper nor a lower bound on the complexity of formulas which can define Henkin-complete extensions of T. For on the one hand, Putnam (1965) showed that it is possible to find such a definition in the class Σ^*_1 consisting of truth functional combinations of Σ^0_1-formulas.[63] And on the other, as long as T is a sound theory it is also possible to increase the complexity of $\gamma(x)$ by modifying parameter vii) (i.e. the extension condition) so that all sentences which are added to Γ fall under the partial truth predicate $\mathrm{Tr}_{\Sigma_{k+1}}(x)$ (which provably cannot be defined by a Π^0_k-formula).
3) The property $\mathrm{Hcm}_{\mathsf{T}}(x)$ which holds of the code of a formula just in case it defines a Henkin-complete extension of T is itself arithmetically definable. But it is easy to see that even if we restrict attention to Δ^0_2-completions, the problem of deciding $\mathrm{Hcm}_{\mathsf{T}}(x)$ is Π^0_2-complete and thus undecidable.[64]
4) Building on Proposition 5.2, it is also possible to define $\gamma(x)$ so that λ_γ is not only independent of T but such that $\mathsf{T} + \gamma$ and $\mathsf{T} + \neg\gamma$ are both interpretable in T — i.e. so that λ_γ is an *Orey sentence* for T.[65]

These observations would appear to rule out a number of otherwise plausible means of imposing conditions on a "canonical" definition of $\gamma(x)$ so as to determine the truth value of λ_γ.[66] This in turn suggests that sentences like λ_γ do not fall naturally into any of Kreisel's three categories of undecidable statements. For in regard to the first, note that sentences like λ_γ are clearly metamathematical in character — e.g. in virtue of their explicit reliance on an arithmetization of syntax. But such sentences are neither Π^0_1 nor are they defined such that the metatheoretic property which they express is clearly true or false on the intended interpretation of either T or its metatheory.

---

[63] Putnam (1957) had previously shown that this cannot be improved to $\Sigma^0_1 \cup \Pi^0_1$. A contemporary refinement of the positive result is the so-called *Low Arithmetized Completeness Theorem* (Hájek & Pudlák, 1998, §I.4b) which provides a more precise syntactic classification in terms of the class $\Sigma^*_0(\Sigma_1)$.

[64] Note on the one hand that the conditions expressing that $\varphi(x)$ defines a Henkin-complete extension of T are Π^0_1 in the definition of $\varphi(x)$. On the other, it is easy to see that there is a reduction from the canonical Π^0_2-complete set TOT of indices of total recursive functions to $\mathrm{Hcm}(x)$.

[65] This fact is mentioned in passing and without proof by Visser (1998, p. 323). One way in which it can be shown is by modifying parameter iii) — i.e. the formalization of provability – in the style of Feferman (1960) and then invoking the Orey-Gauspari-Lindström Theorem (see Lindström, 1997, p. 81).The relevant techniques and several generalizations will be developed in a sequel to this paper.

[66] For instance in light of 1) and 2), there appears to be no principled means of restricting attention to a particular formula class in attempting to define a completion which better approximates TA. And in light of 3), we cannot even uniformly decide which arithmetical formulas we should consider for performing this task.

In regard to Kreisel's second category, note that however $\gamma(x)$ is defined, λ_γ will be provably equivalent to a first-order arithmetical statement. It is thus unlike the Continuum Hypothesis in not admitting a higher-order formulation whose truth value can be affected by the external interpretation of the second-order quantifiers. But at the same time, there are also choices of $\gamma(x)$ for which λ_γ will share an important affinity with CH which, it will be recalled, is an Orey sentence for ZFC. And thus if $\gamma(x)$ is chosen per 4) so that λ_γ is also an Orey sentence for T, then it will be like CH in the sense that the adjunction of either it or its negation to T yields an extension which does not increase consistency strength.

In regard to Kreisel's third category, the received view to which he appears to be reacting in (1967) holds that the choice between Euclidean and non-Euclidean geometries — and thus the truth or falsity of the Parallel Postulate — is *conventional*. On the other hand, λ_γ is an *arithmetical* formula and the received view here holds that arithmetic is unlike geometry in having a determinant subject matter. Proponents of this position will thus hold that once $\gamma(x)$ is fixed, λ_γ will have a determinate truth value in $\mathcal{N}$. On this basis it might initially seem far fetched to think that the choice between λ_γ and $\neg\lambda_\gamma$ could in any sense be conventional.[67]

But to reach such a conclusion summarily would be to overlook the origin of λ_γ and the other undecidable statements which we have been considering. For since $\gamma(x)$ defines a Henkin-complete extension of T, we have seen that it may be understood semantically as defining a model $\mathcal{M} \models \mathsf{T}$ from the standpoint of the standard model $\mathcal{N} \models \mathsf{T} + \mathrm{Con}(\mathsf{T})$ which, on the received view, is the intended interpretation of T. Since λ_γ is derived by diagonalizing $\gamma(x)$, it also seems reasonable to maintain that the description of such a model contributes to its meaning. But as we have seen, not only must any model defined in this way be nonstandard, Proposition 5.2 also demonstrates that we have considerable flexibility in constructing $\mathcal{M}$ — inclusive of controlling whether it satisfies λ_γ itself. Thus absent principled criteria by which we may rule out various candidate definitions of $\gamma(x)$, it would seem that the truth value of the corresponding sentences λ_γ is indeed conventional to at least the extent that the choice between different arithmetically definable nonstandard models of T is.[68]

5.5. Comparison with axiomatic theories of truth A satisfactory assessment of Kreisel and Wang's approach described at the beginning of this section would also need to further integrate the foregoing observations about undecidable statements obtained in §4 with an account of the relationship which they bear to the paradoxes which

---

[67] For instance Koellner (2009) appears to suggest that being an Orey sentence for ZFC is a necessary but not sufficient condition for regarding the truth value of a set theoretical statement as one of "mere expedience". On the other hand, he rules out the possibility that arithmetical Orey sentences have this status by fiat.

[68] An analogous sequence of observations will be true for the other undecidable statements constructed in §4. For note that the only features of the sentence λ_γ on which Proposition 5.2 depends are that it is undecidable in T and that it may be constructed effectively from the definition of $\gamma(x)$. A version of this result will thus be true, e.g., for the Kreisel sentence $k \in K$ constructed in the proof of Proposition 4.3. But now suppose that we take $\mathcal{M} \models \mathsf{S} + \mathrm{Con}(\mathsf{S})$ to be the intended model of S. Then it will be possible to define predicates $\gamma_0(x)$ and $\gamma_1(x)$ defining Henkin-complete extensions of S such that the sentences $k_0 \in K_0$ and $k_1 \in K_1$ obtained from the corresponding interpretations $(\cdot)^0$ and $(\cdot)^1$ in the manner of Proposition 4.3 are respectively false and true in $\mathcal{M}$. But via a series of observations analogous to those recorded in §3.3, it is easy to see that $\gamma_0(x)$ and $\gamma_1(x)$ will both define ω-nonstandard extensions of $\mathcal{M}$. It would thus again appear that we are free to choose between them — and also the truth value of the relevant instantiation of the Kreisel sentence — without contravening our conviction that $\mathcal{M}$ is the intended model of S.

are used to generate them. We have already seen in §5.1 that one potential view is that these statements should be regarded as the images of the paradoxical ones under Hilbert & Bernays's method of arithmetization. And thus at least from the perspective of their *finiten Standpunkt*, it is not unreasonable to think that the paradoxes are indeed dissolved in virtue of the relevant undecidability results.

A further comparison can also be framed between Hilbert & Bernays's approach to the semantic paradoxes and that provided by contemporary axiomatic theories of truth. Theories of this sort are typically put forth in an attempt to formalize portions of our reasoning about the "everyday" notion of truth or *truth simpliciter* as it is sometimes called — i.e. truth in some fixed intended interpretation rather than in an arbitrary model.[69] As such, an axiomatic theory of truth is typically formed by adjoining to PA axioms or rules involving a *primitive* predicate $\mathtt{T}(x)$ intended to express this notion. Examples of such $\mathcal{L}_{\mathtt{T}} = \mathcal{L}_{\mathsf{Z}} \cup \{\mathtt{T}(x)\}$ theories include PAT, CT, FS, and KF, familiarity with which will be assumed in the sequel.[70]

The properties of $\mathtt{T}(x)$ which are axiomatized by these theories may be contrasted with those which hold for the various *defined* notions of truth we have been considering. Of these, it seems reasonable to take as basic the $\mathcal{L}_{\mathsf{Z}}^2$ truth definition $\mathrm{Tr}(x)$ introduced in §4.3.1. As we have seen, this definition essentially involves impredicative second-order quantification. Hilbert & Bernays thus possessed a natural motivation to interpret $\mathrm{Tr}(x)$ via the method of arithmetization. This may in turn may be achieved by applying the Arithmetized Completeness Theorem to obtain a mapping $(\cdot)^* : \mathcal{L}_{\mathsf{Z}}^2 \to \mathcal{L}_{\mathsf{Z}}$ which is an interpretation of a theory such as ACA_0 in which it is possible to develop the properties of $\mathrm{Tr}(x)$ relative to a first-order theory such as $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0)$.

One axis along which it is possible to compare the properties of these formulas with those of a primitive truth predicate $\mathtt{T}(x)$ is with respect to the status of the resulting Liar-like sentences. Recall that the sentences λ_{Tr} and λ_{Tr^*} are respectively provable and undecidable over ACA_0. To assess the status of the analogous sentence for $\mathtt{T}(x)$ let $\lambda_{\mathtt{T}}$ be $\lambda_{\mathtt{T}} =_{\mathrm{df}} \neg\mathtt{T}(\mathrm{subst}(\bar{l}, \bar{l}))$ where $l = \ulcorner \mathtt{T}(\mathrm{subst}(x, x)) \urcorner$. This will be a Δ_0^0-formula of $\mathcal{L}_{\mathtt{T}}$ if we assume that the substitution function is in the language. But although in this case we have $\mathsf{PAT} \vdash \lambda_{\mathtt{T}} \leftrightarrow \neg\mathtt{T}(\ulcorner \lambda_{\mathtt{T}} \urcorner)$, it is a consequence of well-known results that the status of $\lambda_{\mathtt{T}}$ depends on the truth-theoretic axioms which are adopted.

PROPOSITION 5.3 i) $\lambda_{\mathtt{T}}$ *is derivable in any theory extending* PAT *which includes the axiom* $\forall x(\mathtt{T}(x) \to \mathrm{Sent}_{\mathcal{L}_{\mathsf{Z}}}(x))$ *asserting that only arithmetical sentences fall under* $\mathtt{T}(x)$; ii) $\lambda_{\mathtt{T}}$ *is also derivable in* KF *presuming that this theory is understood to include the consistency axiom* CONS (*i.e.* $\forall x(\mathrm{Sent}_{\mathcal{L}_{\mathtt{T}}}(x) \to \neg(\mathtt{T}(x) \wedge \mathtt{T}(\dot{\neg}x))))$; iii) $\lambda_{\mathtt{T}}$ *is undecidable in* FS *presuming that* FS *is consistent.*

Although the axiom $\forall x(\mathtt{T}(x) \to \mathrm{Sent}_{\mathcal{L}_{\mathsf{Z}}}(x))$ is not typically included in any of the theories mentioned above, it may be naturally adjoined to theories such as CT which are intended to axiomatize a *typed* notion of truth in the sense that their axioms do not imply that $\mathtt{T}(x)$ holds of any statements which contain this predicate itself. As such, part i) of Proposition 5.3 reflects that if we took the additional step of asserting that $\mathtt{T}(x)$ *cannot* hold of such statements, then we would be able to prove $\neg\mathtt{Tr}(\ulcorner \lambda_{\mathtt{T}} \urcorner)$ — and hence also $\lambda_{\mathtt{T}}$ — in virtue of the fact that $\lambda_{\mathtt{T}}$ is an $\mathcal{L}_{\mathtt{T}}$-sentence which is not an $\mathcal{L}_{\mathsf{Z}}$-sentence. This can be compared to the fact that λ_{Tr} is provable in ACA_0 in virtue of the fact that it is an $\mathcal{L}_{\mathsf{Z}}^2$-sentence which is also not an $\mathcal{L}_{\mathsf{Z}}$-sentence.

On the other hand, no such restriction is imposed by the untyped theories FS and KF, both of which contain principles which imply that $\mathtt{T}(x)$ holds of sentences containing this predicate. In this setting the undecidability of $\lambda_{\mathtt{T}}$ is sometimes presented as a *positive*

[69] See, e.g., (Dummett, 1978, p. 123), (Field, 2008, §1), and (Cieśliński, 2018, §2).
[70] For axiomatizations and accounts of the results assumed below see (Halbach, 2011).

feature of FS as it is taken to accord with the intuition that the Liar sentence obtained from a predicate formalizing the "everyday" notion of truth does not have a determinate truth value.[71] On the other hand, the derivability of $\lambda_{\mathtt{Tr}}$ – and thus also $\neg\mathtt{T}(\ulcorner\lambda_{\mathtt{Tr}}\urcorner)$ – has been presented as evidence that KF is not a "natural" or "intuitively plausible" theory of truth.[72]

We have seen that an interpretation $(\cdot)^* : \mathcal{L}_{\mathsf{Z}}^2 \to \mathcal{L}_{\mathsf{Z}}$ may be understood as a type-lowering operator in the sense that the image of an $\mathcal{L}_{\mathsf{Z}}^2$-sentence under $(\cdot)^*$ is an $\mathcal{L}_{\mathsf{Z}}$-sentence. From this it follows that $\mathrm{Tr}^*(\ulcorner\varphi\urcorner)$ will not always be refutable in ACA_0 in instances in which φ contains the formula $\mathrm{Tr}^*(x)$ itself. For instance, we have seen that if $\mathsf{PA} \vdash \varphi$, then $\mathsf{PA} \vdash \mathrm{Tr}^*(\ulcorner\mathrm{Tr}^*(\ulcorner\varphi\urcorner)\urcorner)$ and we have also seen that λ_{Tr^*} is undecidable even in ACA_0 (presuming that $\mathsf{ACA}_0 + \mathrm{Con}(\mathsf{ACA}_0)$ is consistent).

The latter result is analogous to the undecidability of $\lambda_{\mathtt{T}}$ in FS in that both results can be obtained by noting that $\mathrm{Tr}^*(x)$ and $\mathtt{T}(x)$ both satisfy necessitation and consistency-cum-completeness conditions similar to (5.6a,b). But since $\mathtt{T}(x)$ is a primitive predicate, it seems that Proposition 5.3iii) does little more than reaffirm a brute intuition about the status of the Liar.[73] On the other hand, the considerations adduced in §5.4 suggest an explanation of *why* it is reasonable to regard λ_{Tr^*} as indeterminate in truth value in virtue of the intensional exigencies which must be confronted to define a predicate which provably satisfies these properties.

These considerations suggest that it may be of independent interest to explore the use of arithmetically definable predicates such as $\mathrm{Tr}^*(x)$ to obtain models of various axiomatic theories of truth. If we take to heart Hilbert & Bernays's admonishments about the foundational significance of the truth predicate (1939, pp. 338-339/351-353), we should presumably view truth-theoretic reasoning as subsumed by a portion of mathematics whose consistency must be accounted for by other means.[74] The natural targets in this regard thus become theories such as $\mathsf{CT}\restriction$ and $\mathsf{FS}\restriction$ — i.e. the typed compositional and Friedman-Sheard theories with induction restricted to $\mathcal{L}_{\mathsf{Z}}$ — which are conservative over PA.

One feature which is often cited in favor of these systems is that they validate a disquotational characterization of truth in that they prove the T-biconditionals for all $\mathcal{L}_{\mathsf{Z}}$-sentences. On the other hand, since $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0) \vdash \varphi^* \leftrightarrow \gamma(\ulcorner\varphi\urcorner)$ for all $\mathcal{L}_{\mathsf{Z}}^2$-sentences, it additionally follows from the provability of the T-biconditionals (4.16) in ACA_0 that

$$\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0) \vdash \varphi^* \leftrightarrow \mathrm{Tr}^*(\ulcorner\varphi\urcorner) \tag{5.8}$$

for all $\mathcal{L}_{\mathsf{Z}}$-sentences φ. This illustrates a sense in which the arithmetical predicate $\mathrm{Tr}^*(x)$ also satisfies something akin to a disquotation scheme. For suppose $\mathcal{M} \models \mathsf{ACA}_0$ is strongly defined by $(\cdot)^*$ relative to the standard model $\mathcal{N}$ in the manner of Theorem 3.4. As the way in which the non-logical symbols of φ are interpreted in $\mathcal{M}$ mirror the substitution of arithmetical formulas induced by $(\cdot)^*$, (5.8) illustrates how $\mathrm{Tr}^*(x)$ may be viewed as a definition of truth for $\mathcal{M}$. In particular, we have that for all $\mathcal{L}_{\mathsf{Z}}$-sentences φ

$$\mathcal{N} \models \mathrm{Tr}^*(\ulcorner\varphi\urcorner) \text{ if and only if } \mathcal{M} \models \varphi \tag{5.9}$$

---

[71] On this point see (Kripke, 1975, p. 707, pp. 714-715), (Reinhardt, 1986, p. 238), (McGee, 1990, p. 14), (Feferman, 1991, p. 41).

[72] See, e.g., (Halbach & Horsten, 2006, p. 682). This paper also presents a "partial" version of KF without CONS in which the undecidability of $\lambda_{\mathtt{T}}$ is restored.

[73] See, e.g., (Horsten, 2011, p. 112).

[74] It would be anachronistic to try to locate the views of either Hilbert or Bernays with respect to the latter day debate about deflationism. But it may still be noted that in (1939, §5.2e) they appear to disavow the instrumental use of truth-theoretic reasoning in consistency proofs of the sort considered by Reinhardt (1986) or Feferman (1991).

As we have seen in §4.1, the first-order part of any model $\mathcal{M}$ which has been constructed in this manner will be a non-elementary end extension of $\mathcal{N}$ – e.g. we know in virtue of Proposition 4.5 that these models must differ in the truth value they assign to λ_{Tr^*}. On the other hand, we have also seen in §5.4 that there is no evident way of using general principles about truth — e.g. of the sort underlying the inductive argument considered in §5.1 — to decide the truth of such sentences. And thus since $\mathcal{M}$ presumably satisfies all of the statements in which we might have a prior *mathematical* interest, the fact that it extends $\mathcal{N}$ non-elementarily for some sentences which can be contrived metamathematically need not be taken to tell against the potential use of $\mathrm{Tr}^*(x)$ to explain various other aspects of our reasoning about truth.

Note, for instance, that since the Tarskian compositional clauses for $\mathrm{Tr}(x)$ are provable in ACA_0, we additionally have (e.g.) $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0) \vdash \neg\varphi^* \leftrightarrow \mathrm{Tr}^*(\ulcorner\neg\varphi\urcorner)$ and similarly for the clauses for the other connectives. Since $(\cdot)^*$ is an interpretation rather than a predicate, this can only be asserted schematically. But as we have seen in §3.3, the formula $\gamma(x)$ will provably satisfy uniform analogs of the Tarskian compositional clauses in virtue of (3.7c,d,e). This can be used to show that $\mathrm{Tr}^*(x)$ and $\gamma(x)$ define coextensive notions of truth for arithmetical sentences in the sense that $\mathsf{ACA}_0 \vdash \forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \rightarrow (\mathrm{Tr}^*(x) \leftrightarrow \gamma(x)))$.[75] Combining these facts allows us to show that over $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0)$ we can derive uniform versions of the compositional clauses such as

$$\forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \rightarrow (\mathrm{Tr}^*(\dot{\neg}x) \leftrightarrow \neg\mathrm{Tr}^*(x))) \tag{5.10}$$

and thus also $\mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0) \vdash \forall x(\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x) \rightarrow (\mathrm{Tr}^*(\dot{\neg}x) \vee \mathrm{Tr}^*(x)))$.

This appears to show that by employing $\mathrm{Tr}^*(x)$, we are able to satisfy the condition discussed in §5.1 which Tarski ultimately imposed on an adequate theory of truth. However, a caveat is again necessary in virtue of the fact that $\mathrm{Tr}^*(x)$ defines truth in $\mathcal{M}$ rather than in $\mathcal{N}$. A consequence is that the truth of (5.10) in $\mathcal{N}$ only entails that negation commutes with $\mathrm{Tr}^*(x)$ for *standard length* $\mathcal{L}_\mathsf{Z}$-sentences rather than those which fall under the interpretation of $\mathrm{Sent}_{\mathcal{L}_\mathsf{Z}}(x)$ in $\mathcal{M}$ (which, by a familiar overspill argument, will contain the codes of statements of nonstandard length). More generally, if the Henkin construction described by $\gamma(x)$ is carried out in an arbitrary model of $\mathcal{M}_1 \models \mathsf{PA} + \mathrm{Con}(\mathsf{ACA}_0)$ to construct an end extension $\mathcal{M}_2 \models \mathsf{ACA}_0$, there is no guarantee that $\{a \in |\mathcal{M}_1| : \mathcal{M}_1 \models \mathrm{Tr}^*(\overline{a})\}$ will be a full satisfaction class for the reduct of $\mathcal{M}_2$ to $\mathcal{L}_\mathsf{Z}$ when defined from within $\mathcal{M}_2$. For unless additional conditions are imposed on $\gamma(x)$, if $\mathcal{M}_2$ is constructed over the standard model, then its reduct to $\mathcal{L}_\mathsf{Z}$ need not be recursively saturated. And in this case it will not admit a satisfaction class by the well-known theorem of Lachlan (1981).

It is, however, possible to use the techniques we have been considering to ensure that a model constructed by the formalized Henkin procedure is recursively saturated. One means of doing this is to take advantage of the following fact:[76]

PROPOSITION 5.4 *Suppose that $\mathcal{M}_1 \models \mathsf{PA}$ is nonstandard and $\mathcal{M}_2$ is strongly definable in $\mathcal{M}_1$. Then $\mathcal{M}_2$ is recursively saturated.*

As we have seen, if $\mathcal{M}_2 \models \mathsf{PA}$ is strongly defined in $\mathcal{M}_1 \models \mathsf{PA} + \mathrm{Con}(\mathsf{PA})$ by a formula $\gamma(x)$ defining a Henkin-completion, then $\mathcal{M}_2$ must be nonstandard. Since $\mathrm{Con}(\mathsf{PA})$ is undecidable in PA, it is possible to define $\gamma(x)$ so that $\mathcal{M}_2 \models \mathrm{Con}(\mathsf{PA})$ — e.g. by placing $\mathrm{Con}(\mathsf{PA})$ first in the enumeration on which the definition of $\gamma(x)$ is based as discussed in §5.4. In this case, the construction described by $\gamma(x)$ can again be carried out in $\mathcal{M}_2$

[75] Since $\mathrm{Tr}^*(x)$ is an $\mathcal{L}_\mathsf{Z}$-formula, this may be shown by an internal induction on the complexity of formulas in ACA_0 making use of the fact that both this predicate and $\gamma(x)$ satisfy the compositional clauses.

[76] See, e.g., (Smorynski, 1984, §6).

yielding a model $\mathcal{M}_3 \models \mathsf{PA}$ which must be recursively saturated by Proposition 5.4. But now note that the iterated procedure is already describable from the perspective of $\mathcal{M}_1$ by considering the formula $\gamma^2(x) = \gamma(\ulcorner\gamma(\dot{x})\urcorner)$.

It is a consequence of another well-known theorem of Kotlarski, Krajewski, & Lachlan (1981) that any countable recursively saturated model $\mathcal{M} \models \mathsf{PA}$ admits a satisfaction class $S \subseteq |\mathcal{M}|$. From this it follows that $\langle \mathcal{M}, S \rangle$ will be a model of $\mathsf{CT}\!\uparrow$. Three questions which arise naturally at this stage but which will have to await other occasions are as follows: 1) Can the definition of such a satisfaction class be formalized in PA (or a weaker theory) by generalizing the techniques of §3.3? 2) If so, can the iteration of this method be formalized to obtain a definition of a *type-free satisfaction class*[77] which would provide an interpretation of $\mathsf{FS}\!\uparrow$ in PA? 3) How would such constructions interact with the other intensional parameters which have been suggested in §5.4 to be intrinsic to the arithmetization of the completeness theorem?[78]

§6. Conclusion The goal of this paper has been to offer a systematic reconstruction and generalization of a method for transforming paradoxes into formal incompleteness results originally due to Kreisel and Wang, building on work of Hilbert and Bernays. The following is a summary of the historical points which I have sought to highlight and synthesize:

1) Hilbert & Bernays's presentation of Gödel's incompleteness theorems in the second volume of the *Grundlagen der Mathematik* (1939) was not isolated from the other aspects of their development of the finitary standpoint. Rather they refine and elaborate on prior discussions of the set theoretic and semantic paradoxes, their development of second-order logic, and their appreciation of the diagonal method — e.g. in (Hilbert, 1917), (Hilbert & Ackermann, 1928), and (Hilbert & Bernays, 1934, §7). On this basis, they also made substantial advances to the formalization of semantic concepts leading to their formal truth definition for first-order arithmetic presented in (1939, §5.2e).
2) Bernays appreciated the relationship between these developments and Gödel's (1930) proof of the completeness theorem for first-order logic, which itself was grounded in concepts and methods presented in these prior sources. In particular, he realized Gödel's proof could be formalized in an arithmetical language, which in turn led to his formulation of the Arithmetized Completeness Theorem in (1939, §4.2).
3) This result can also be understood as a mathematical embodiment of what Hilbert & Bernays refer to as the *method of arithmetization* (1934, pp. 1-3, 18-19) which they (and others in the Göttingen school) had sought to employ in consistency proofs for

[77] See (Halbach, 2011, §14.1).

[78] A positive answer to question 1) is stated by Enayat & Visser (2015, §16.5) on the basis of their method for constructing full satisfaction classes which avoids the techniques of M-logic originally employed by Kotlarski, Krajewski, & Lachlan (1981). In fact by using either a formalized version of this construction or the arithmetization of the conservativity proof for $\mathsf{CT}\!\uparrow$ via the corrected cut elimination argument of Leigh (2015), it is in principle possible to obtain an interpretation of $\mathsf{CT}\!\uparrow$ in PA (as was originally observed by Fischer, 2009). The full details of the relevant techniques have not yet been published. But it may still be noted in regard to question 3) that they all rely on variants of the Arithmetized Completeness Theorem — e.g. in order to formalize the compactness and elementary chain arguments of (Enayat & Visser, 2015) or to obtain an interpretation on the basis of the conservativity of $\mathsf{CT}\!\uparrow$ via the Orey Compactness Theorem (see, e.g., Lindström, 1997, p. 80).

instances of "formal axiomatics" since the 1890s. This included axiomatizations not only of Euclidean and non-Euclidean geometries, but also of mathematical physics, analysis, and set theory which Hilbert & Bernays speak of as "transcend[ing] the realm of experience and intuitive self-evidence" (1934, p. 3/3).

4) Kreisel (1950, 1953) described a means of combining Bernays's formalization of Gödel's completeness theorem and his incompleteness theorems by applying the former to an axiomatic system S of sets and classes similar to Gödel-Bernays set theory. In this way he showed that it is possible to formulate a sentence κ akin to the statement that the Russell class is not a member of itself. Although κ is refutable in S, Kreisel showed that its image κ^* under an interpretation derived from the Arithmetized Completeness Theorem is formally undecidable in S (provided that $\mathsf{S} + \mathrm{Con}(\mathsf{S})$ is consistent). Kreisel's method was later presented in a more accessible form by Wang (1955). Wang also generalized Kreisel's treatment of Russell's paradox to the semantic paradoxes by showing how Hilbert & Bernays's truth definition can be used to produce a Liar-like sentence whose image under $(\cdot)^*$ is also formally undecidable.

Wang's result is in turn similar to Bernays's original explanation of how arithmetical formalization transforms the Liar paradox into a non-derivability result (1939, §5.2c). As Bernays notes, this renders the system F in which Hilbert & Bernays carried out their formalization incomplete rather than inconsistent. The incompleteness theorems have, of course, persistently been regarded as undermining the goal of the Hilbert program understood as that of offering a finitary consistency proof for infinitary mathematics in which Hilbert and his collaborators had hoped to formalize analysis (and of which systems like GB and ACA_0 are natural exemplars). But it is also notable that Bernays appears to have come to view the results in question as playing a salutary role in his mature understanding of the program. For not only did he suggest that the reception of the incompleteness theorems within proof theory has led to other "constructive" techniques for proving consistency (e.g. Bernays, 1954b, pp. 10-11), he went to considerable lengths to illustrate how the method of arithmetization itself leads to such a proof for a fragment of his own axiomatization of set theory (Bernays, 1954a, §19). For as we have seen, such a construction can be understood as showing that even if we work over a language which is sufficiently expressive to formulate the relevant "paradoxical notions", the corresponding "paradoxical statements" do not give rise to contradictions but are rather transformed into undecidable ones.

The events just summarized have their origins within the foundational developments of the 1900s–1930s which have been widely scrutinized by philosophers of mathematics. Nonetheless, the methods and results just described have received little notice outside the developments within mathematical logic to which they directly contributed (on which see Dean, 2017 and Dean & Walsh, 2017). Two apparent reasons for this are as follows. First, despite its classical origins, the Arithmetized Completeness Theorem remains largely unknown to philosophers. Second, Kreisel and Wang's methodological reflections notwithstanding, the possibility that formal incompleteness might contribute to a uniform resolution to the paradoxes (rather than merely being a byproduct of them) has been largely overlooked within latter day work in philosophical logic. I will conclude by offering a brief comment on the prospects for further developments in each of these regards.

As I have attempted to illustrate in §3 and §4, the methods employed in the proof of the Arithmetized Completeness Theorem are both general in scope and flexible with respect to the sorts of interpretations they can be used to construct. In particular, while Theorems 3.2 and 3.4 are traditionally presented as applying to *first-order* theories, it is striking that Kreisel and Wang's original application of these results was to a theory similar to one which is now often described as *second-order* — i.e. Gödel-Bernays set theory. As we have seen, there is no tension here because it is possible to regard both

GB and similar $\mathcal{L}^2_{\mathsf{Z}}$-theories such as ACA_0 as two-sorted first-order systems. But as we are now in a better position to appreciate, when we apply Theorem 3.2 to such a theory to yield an interpretation $\in^*$ of the symbol $\in$ in the language of first-order arithmetic, we are explicitly treating membership as a *non-logical relation* whose extension we can control in the manner illustrated by results like Proposition 5.2. This in turn highlights in vivid terms what is at issue in applying the method of arithmetization to theories sufficient for the development of analysis or set theory.[79]

In §5 I have attempted to lay out some of the philosophical exigencies which Wang's second alternative — i.e. that of "admitting paradoxical classes" but "treating as undecided" questions about membership in them – would presumably have to face were it to be developed systematically. And although I have not articulated an explicit thesis about the resolution of the paradoxes here, we are now also in an improved position to appreciate two potentially attractive features of such an approach. First, we have seen how Wang's proposal grows naturally out of considerations which have their origin in Hilbert & Bernays's finitary standpoint. This in turn suggests that it may be useful to adopt such a perspective in our approach to the paradoxes — e.g. to provide a set of foundational and methodological desiderata against which to compare proposals of the sort which is often lacking in contemporary philosophical theorizing. Second, I have suggested in §5 that the resulting developments can be understood as providing at least a technical explanation of *why* the resulting formalized versions of paradoxical statements are formally undecidable — a feature which has been repeatedly advertised as a *virtue* of various axiomatic approaches to the semantic paradoxes.[80] As I have suggested, however, the ultimate evaluation of Wang's proposal must await a more thorough account of the sort of incompleteness which is at issue – e.g. Is the basis for the undecidability of the statements constructed in §4 best understood in semantic, epistemic, or conventional terms? How do the various dimensions of intensionality discussed in §5 bear on this?

Acknowledgements This paper was originally presented at the conference "Intuitionism, Computation, and Proof: Selected themes from the research of G. Kreisel" held at the Institute for History and Philosophy of Science and Technology (CRNS/Paris 1/ENS UMR 8590) in June 2016. Special thanks are owed to the organizers (Marianna Antonutti Mafori and Mattia Petrolo), the other participants (Michael Detlfesen, Daniel Isaacson, Ulrich Kohlenbach, Angus MacIntyre, Joan Moschovakis, David McCarty, Dana Scott, Göran Sundholm, and Mark van Atten) and to Volker Halbach, Richard Kaye, Hidenori Kurokawa, Benedict Eastaugh, and Sean Walsh for comments. Thanks are also owed to audiences at Birmingham, Oxford, Notre Dame, Irvine, and Saint Andrews where this material was presented. I am also indebted to a particularly scrupulous and helpful anonymous referee.

Bibliography

Page references to Hilbert & Bernays, *Grundlagen der Mathematik*, Volumes I and II are in the form p. m/n for m the page in the first edition and n the page in the second edition.

---

[79] This too is a point which Bernays would later acknowledge (e.g. 1970, p. 184). It additionally foretells of the possibility of formalizing subsequent independence proofs in set theory by refinements to the Henkin construction – a point to which Cohen (1966, p. 112) would himself later allude.

[80] For a recent example see (Fischer, Horsten, & Nicolai, 2019). But as in the case of Fitch's (1952) proposal discussed in note 54, the "silence" of the truth theory proposed therein with respect to the Liar is achieved by fiat via the adoption of a non-classical background logic.

Ackermann, W. (1928). Über die Erfüllbarkeit gewisser Zählausdrücke. *Mathematische Annalen* **100**(1), 638–649.

Bernays, P. (1930). Die Philosophie der Mathematik und die Hilbertsche Beweistheorie. *Blätter für deutsche Philosophie* **4**, 326–367. Reprinted in Bernays (1976), pp. 17–62 and in Mancosu (1998), pp. 234-265.

Bernays, P. (1937). A system of axiomatic set theory: Part I. *Journal of Symbolic Logic* **2**(1), 65–77.

Bernays, P. (1942). A System of Axiomatic Set Theory: Part III. Infinity and Enumerability. Analysis. *Journal of Symbolic Logic* **7**(2), 65–89.

Bernays, P. (1954a). A system of axiomatic set theory: Part VII. *Journal of Symbolic Logic* **19**(2), 81–96.

Bernays, P. (1954b). Zur Beurteilung der Situation in der beweistheoretischen Forschung. *Revue Internationale de Philosophie* **27/28**, 9–13.

Bernays, P. (1970). Die schematische Korrespondenz und die idealisierten Strukturen. *Dialectica* **24**, 53–66. Reprinted in Bernays (1976), pp. 176–188.

Bernays, P. (1976). *Abhandlungen zur Philosophie der Mathematik*. Darmstadt: Wiss. Buchgesellschaft.

Bernays, P., & Fraenkel, A. (1958). *Axiomatic Set Theory*. Amsterdam: North-Holland.

Boolos, G. (1989). A new proof of the Gödel incompleteness theorem. *Notices of the American Mathematical Society* **36**(4), 388–390.

Borel, É. (1898). *Leçons sur la théorie des fonctions*. Paris: Gauthier-Villars et fils.

Church, A. (1956). *Introduction to mathematical logic*. Princeton: Princeton University Press.

Church, A. (1976). Comparison of Russell's resolution of the semantical antinomies with that of Tarski. *The Journal of Symbolic Logic* **41**(4), 747–760.

Cieśliński, C. (2002). Heterologicality and incompleteness. *Mathematical Logic Quarterly* **48**(1), 105–110.

Cieśliński, C. (2018). *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge: Cambridge University Press.

Cohen, P. (1966). *Set theory and the continuum hypothesis*. New York: W.A. Benjamin.

Dean, W. (2015). Arithmetical reflection and the provability of soundness. *Philosophia Mathematica* **23**(1), 31–64.

Dean, W. (2017). Bernays and the Completeness Theorem. *Annals of the Japanese Association for the Philosophy of Science* **25**, 44–55.

Dean, W., & Walsh, S. (2017). The prehistory of the subsystems of second-order arithmetic. *The Review of Symbolic Logic* **10**(2), 357–396.

Doets, K. (1999). Relatives of the Russell Paradox. *Mathematical Logic Quarterly* **45**, 73–83.

Dummett, M. (1978). Frege's distinction between sense and reference. In *Truth and other enigmas*, pp. 116–144. Cambridge, MA: Harvard University Press.

Ebbs, G. (2015). Satisfying predicates: Kleene's proof of the Hilbert–Bernays Theorem. *History and Philosophy of Logic* **36**(4), 346–366.

Enayat, A., & Visser, A. (2015). New constructions of satisfaction classes. In Achourioti, T., Galinon, H., Fernández, J. M., & Fujimoto, K., editors, *Unifying the Philosophy of Truth*, pp. 321–335. Dordrecht: Springer.

Ewald, W. (1996). *From Kant to Hilbert: A Source Book in the Foundations of Mathematics*. New York: Oxford University Press.

Ewald, W., & Sieg, W., editors (2013). *David Hilbert's Lectures on the Foundations of Logic and Arithmetic 1917 – 1933*. Berlin: Springer.

Feferman, S. (1960). Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae* **49**, 35–92.

Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic* **56**(1), 1–49.

Feferman, S., et al., editors (1986). *Kurt Gödel Collected Works. Vol. I. Publications 1929-1936*. Oxford: Oxford Univeristy Press.

Feferman, S., et al., editors (2003). *Kurt Gödel Collected Works. Vol. IV. Publications Correspondence A-G*. Oxford: Oxford Univeristy Press.

Field, H. (2008). *Saving Truth from Paradox*. Oxford: Oxford University Press.

Fischer, M. (2009). Minimal truth and intepretability. *The Review of Symbolic Logic* **2**, 799–815.

Fischer, M., Horsten, L., & Nicolai, C. (2019). Hypatia's silence: Truth, justification, and entitlement. Forthcoming in *Noûs*.

Fitch, F. (1952). *Symbolic logic*. New York: The Ronald Press Company.

Gentzen, G. (1936). Die Widerspruchsfreiheit der reinen Zahlentheorie. *Mathematische Annalen* **112**, 493–565.

Gödel, K. (1930). The completeness of the axioms of the functional calculus of logic. pp. 103–123. Reprinted in Feferman et al. (1986).

Gödel, K. (1931a). Correspondence with Ernest Zermelo. Reprinted in Feferman et al. (2003).

Gödel, K. (1931b). On formally undecidable propositions of *Principia Mathematica* and related systems I. Reprinted in Feferman et al. (1986).

Gödel, K. (1940). *The consistency of the axiom of choice and of the generalized continuum-hypothesis with the axioms of set theory*. Princeton: Princeton University Press.

Hájek, P., & Pudlák, P. (1998). *Metamathematics of First-Order Arithmetic*. Berlin: Springer. First edition 1993.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Halbach, V., & Horsten, L. (2006). Axiomatizing Kripke's theory of truth. *The Journal of Symbolic Logic* **71**(2), pp. 677–712.

Halbach, V., & Visser, A. (2014a). Self-reference in arithmetic I. *The Review of Symbolic Logic* **7**(04), 671–691.

Halbach, V., & Visser, A. (2014b). Self-reference in arithmetic II. *The Review of Symbolic Logic* **7**(04), 692–712.

Henkin, L. (1949). The completeness of the first-order functional calculus. *Journal of Symbolic Logic* **14**(03), 159–166.

Hilbert, D. (1899). *Grundlagen der Geometrie*. Leipzig: Teubner.

Hilbert, D. (1916). The Foundations of Physics: The Lectures (1916–1917). Reprinted in (Sauer & Majer, 2009).

Hilbert, D. (1917). Lectures on the principles of mathematics 'prinzipien der mathematik' (ws 1917/18). Reprinted in Ewald & Sieg (2013), pp. 31-274.

Hilbert, D. (1922). Neubegründung der Mathematik: Erste Mitteilung. *Abhandlungen aus dem Seminar der Hamburgischen Universität* **1**, 157–77. English translation as "The new grounding of mathematics: First report" in Ewald (1996), pp. 1115–1134.

Hilbert, D. (1926). Über der Unendliche. *Mathematische Annalen* **95**, 161–190. English translation as "On the infinite" in van Heijenoort (1967), pp. 367-292.

Hilbert, D., & Ackermann, W. (1928). *Grundzüge der theoretischen Logik* (First ed.). Springer. Reprinted in Ewald & Sieg (2013).

Hilbert, D., & Ackermann, W. (1938). *Grundzüge der theoretischen Logik* (Second ed.). Springer. Translated as Hilbert & Ackermann (1950).

Hilbert, D., & Ackermann, W. (1950). *Principles of Mathematical Logic*. New York: Chelsea Publishing Company.

Hilbert, D., & Bernays, P. (1934). *Grundlagen der Mathematik*, Volume I. Berlin: Springer. Second edition 1968.

Hilbert, D., & Bernays, P. (1939). *Grundlagen der Mathematik*, Volume II. Berlin: Springer. Second edition 1970.

Horsten, L. (2011). *The Tarskian Turn: Deflationism and Axiomatic Truth*. Cambridge, MA: MIT Press.

Isaacson, D. (2011). The Reality of Mathematics and the Case of Set Theory. In Novak, Z. & Simonyi, A., editors, *Truth, Reference, and Realism*, pp. 1–75. Budapest: Central European University Press.

Kanamori, A. (2009). Bernays and Set Theory. *Bulletin of Symbolic Logic* **15**(1), 43–69.

Kaye, R. (1991). *Models of Peano Arithmetic*, Volume 15 of *Oxford Logic Guides*. Oxford: Oxford University Press.

Kaye, R., & Wong, T. (2007). On interpretations of arithmetic and set theory. *Notre Dame Journal of Formal Logic* **48**(4), 497–510.

Kikuchi, M. (1997). Kolmogorov complexity and the second incompleteness theorem. *Archive for Mathematical Logic* **36**(6), 437–443.

Kikuchi, M., & Kurahashi, T. (2016). Liar-type paradoxes and the incompleteness phenomena. *Journal of Philosophical Logic* **45**(4), 381–398.

Kikuchi, M., Kurahashi, T., & Sakai, H. (2012). On proofs of the incompleteness theorems based on Berry's paradox by Vopěnka, Chaitin, and Boolos. *Mathematical Logic Quarterly* **58**(4-5), 307–316.

Kikuchi, M., & Tanaka, K. (1994). On formalization of model-theoretic proofs of Gödel's theorems. *Notre Dame Journal of Formal Logic* **35**(3), 403–412.

Kleene, S. (1943). Recursive predicates and quantifiers. *Transactions of the American Mathematical Society* **53**(1), 41–73.

Kleene, S. (1952). *Introduction to Metamathematics*. Amsterdam: North-Holland.

Koellner, P. (2009). Truth in Mathematics: The Question of Pluralism. In Linnebo, O. & Bueno, O., editors, *New Waves in the Philosophy of Mathematics*, pp. 80–116. New York: Palmgrave.

Kotlarski, H. (2004). The incompleteness theorems after 70 years. *Annals of Pure and Applied Logic* **126**(1), 125 – 138.

Kotlarski, H., Krajewski, S., & Lachlan, A. (1981). Construction of satisfaction classes for nonstandard models. *Canadian Mathematical Bulletin* **14**(3), 283–293.

Kreisel, G. (1950). Note on arithmetic models for consistent formulae of the predicate calculus. *Fundamenta mathematicae* **37**, 265–285.

Kreisel, G. (1952). On the concepts of completeness and interpretation of formal systems. *Fundamenta mathematicae* **39**, 103–127.

Kreisel, G. (1953). Note on arithmetic models for consistent formulae of the predicate calculus. II. In *Actes du XIeme Congres International de Philosophie*, Volume XIV, Amsterdam, pp. 39–49. North-Holland.

Kreisel, G. (1955). Models, translations and interpretations. In Skolem, T., editor, *Mathematical interpretation of formal systems*, pp. 26–50. Amsterdam: North Holland.

Kreisel, G. (1958). Wittgenstein's remarks on the foundations of mathematics. *The British Journal for the Philosophy of Science* **9**(34), 135–158.

Kreisel, G. (1965). Mathematical logic. In Saaty, T., editor, *Lectures on Modern Mathematics, Vol. III*, pp. 95–195. New York: Wiley.

Kreisel, G. (1967). Informal Rigour and Completeness Proofs. In *Problems in the Philosophy of Mathematics*, Volume Lakatos, I., pp. 138–186. Amsterdam: North-Holland.

Kreisel, G. (1968). A survey of proof theory. *The Journal of Symbolic Logic* **33**(3), 321–388.

Kreisel, G. (1969). Two Notes on the Foundations of Set-Theory. *Dialectica* **23**(2), 93–114.

Kreisel, G., & Wang, H. (1955). Some applications of formalized consistency proofs. *Fundamenta mathematicae* **42**, 101–110.

Kripke, S. (1975). Outline of a theory of truth. *Journal of Philosophy* **72**(19), 690–716.

Kripke, S. (2014). The Road to Gödel. In Berg, J., editor, *Naming, Necessity, and More*, pp. 223–241. Berlin: Springer.

Kritchman, S., & Raz, R. (2010). The surprise examination paradox and the second incompleteness theorem. *Notices of the AMS* **57**(11), 1454–1458.

Kruse, A. (1963). A method of modelling the formalism of set theory in axiomatic set theory. *The Journal of Symbolic Logic* **28**(1), 20–34.

Lachlan, A. (1981). Full satisfaction and recursive saturation. *Canadian Mathematical Bulletin* **24**(3), 295–297.

Lebesgue, H. (1905). Sur les fonctions représentables analytiquement. *Journal de mathematiques pures et appliquees* **1**, 139–216.

Leigh, G. (2015). Conservativity for theories of compositional truth via cut elimination. *Journal of Symbolic Logic* **80**(3), 845–865.

Lévy, A. (1976). The role of classes in set theory. *Studies in Logic and the Foundations of Mathematics* **84**, 173–215.

Lindström, P. (1997). *Aspects of Incompleteness*, Volume 10 of *Lecture Notes in Logic*. Berlin: Springer.

Lusin, N. (1925). Sur les ensembles non mesurables B et l'emploi de la diagonale Cantor. *CR Acad. Sci. Paris* **181**, 95–96.

Mancosu, P., editor (1998). *From Brouwer to Hilbert: The Debate on the Foundations of Mathematics in the 1920s*. Oxford: Oxford University Press.

Mancosu, P. (2003). The Russellian influence on Hilbert and his school. *Synthese* **137**, 59–101.

Manevitz, L., & Stavi, J. (1980). Operators and alternating sentences in arithmetic. *Journal of Symbolic Logic* **45**(01), 144–154.

McGee, V. (1990). *Truth, Vagueness and Paradox*. Indianapolis: Hackett Publishers.

Mendelson, E. (1997). *Introduction to Mathematical Logic* (Sixth ed.). Boca Raton: CRC Press.

Montague, R. (1955). On the paradox of grounded classes. *Journal of Symbolic Logic* **20**(2), p. 140.

Mostowski, A. (1950). Some impredicative definitions in the axiomatic set-theory. *Fundamenta mathematicae* **38**, 110–124.

Müller, G. (1976). *Sets and classes: On the work by Paul Bernays*, Volume 84 of *Studies in Logic and the Foundations of Mathematics*. Amsterdam: North-Holland.

Myhill, J. (1952). The hypothesis that all classes are nameable. *Proceedings of the National Academy of Sciences* **38**(11), 979–981.

Nelson, L. (1959). *Beiträge zur Philosophie der Logik und Mathematik*. Frankfurt am Main: Öffentliches Leben.

Nelson, L., & Grelling, K. (1908). Bemerkungen zu den Paradoxien von Russell und Burali-Forti. *Abhandlungen der Fries' schen Schule, Neue Folge* **2**, 301–334. Reprinted in Nelson (1959), pp. 57-77.

Novak, I. (1950). A construction for consistent systems. *Fundamenta mathematicae* **1**(37), 87–110.

Peckhaus, V., & Kahle, R. (2002). Hilbert's paradox. *Historia Mathematica* **29**, 99.

Priest, G. (1994). The structure of the paradoxes of self-reference. *Mind* **103**(409), 25–34.

Priest, G. (1997a). On a paradox of Hilbert and Bernays. *Journal of philosophical logic* **26**(1), 45–56.

Priest, G. (1997b). Yablo's paradox. *Analysis* **57**(4), 236–242.

Putnam, H. (1957). Arithmetic models for consistent formulae of quantification theory. *The Journal of Symbolic Logic* **22**, 110–111.

Putnam, H. (1965). Trial and error predicates and the solution to a problem of Mostowski. *Journal of Symbolic Logic* **30**(01), 49–57.

Quine, W. (1981). *Mathematical logic*. Cambridge, MA: Harvard University Press.

Rabin, M. (1958). On recursively enumerable and arithmetic models of set theory. *Journal of Symbolic logic* **23**(4), 408–416.

Ramsey, F. P. (1926). The foundations of mathematics. *Proceedings of the London Mathematical Society* **2**(1), 338–384.

Read, S. (2016, August). Denotation, paradox and multiple meanings. Manuscript.

Reinhardt, W. (1986). Some remarks on extending and interpreting theories with a partial predicate for truth. *Journal of Philosophical Logic* **15**(2), 219–251.
Richard, J. (1905). The principles of mathematics and the problem of sets. Reprinted in van Heijenoort (1967), pp. 142-144.
Robinson, A. (1963). On languages which are based on non-standard arithmetic. *Nagoya Mathematical Journal* **22**, 83–117.
Rogers, H. (1987). *Theory of recursive functions and effective computability*. Cambridge, MA: MIT Press. First edition 1967.
Russell, B. (1903). *The Principles of Mathematics*. Cambridge: Cambridge University Press.
Russell, B. (1908). Mathematical logic as based on the theory of types. *American journal of mathematics* **30**(3), 222–262.
Sauer, T., & Majer, U. (2009). *David Hilbert's lectures on the foundations of physics 1915-1927: relativity, quantum theory and epistemology*. Berlin: Springer.
Shoenfield, J. (1954). A Relative Consistency Proof. *The Journal of Symbolic Logic* **19**, 21–28.
Sieg, W., & Ravaglia, M. (2005). David Hilbert and Paul Bernays, *Grundlagen der Mathematik*, (1934, 1939). In Grattan-Guinness, I., editor, *Landmark Writings in Western Mathematics 1640–1940*, pp. 981. Amsterdam: Elsevier.
Simpson, S. (2009). *Subsystems of second order arithmetic* (second ed.). Cambridge: Cambridge University Press.
Smorynski, C. (1977). The incompleteness theorems. In Barwise, J., editor, *Handbook of Mathematical Logic*, pp. 821–865. Amsterdam: North-Holland.
Smorynski, C. (1984). Lectures on nonstandard models of arithmetic. In Lolli, G., Longo, G., & Marqa, A., editors, *Logic Colloquium '82*, pp. 1–70. Amsterdam: North-Holland.
Smoryński, C. (1985). *Self-reference and modal logic*. Amsterdam: Springer.
Tarski, A. (1935). Der Wahrheitsbegriff in den formalisierten Sprachen. *Studia philosophica* **1**, 261–405. English translation as "The concept of truth in formalized languages" by J.H. Woodger in Tarski (1956).
Tarski, A. (1956). *Logic, Semantics, Metamathematics — Papers from 1923 to 1938*. Oxford: Clarendon Press.
Tarski, A., Mostowski, A., & Robinson, R. (1953). *Undecidable Theories*. Studies in Logic and the Foundations of Mathematics. Amsterdam: North-Holland.
van Heijenoort, J., editor (1967). *From Frege to Gödel : A Source Book in Mathematical Logic, 1879-1931*. Cambridge, MA: Harvard University Press.
Visser, A. (1998). An overview of interpretability logic. In Kracht, M., de Rijke, M., & Wansing, H., editors, *Advances in Modal Logic, Volume 1*, pp. 307–359. Stanford: CSLI Publications.
Von Neumann, J. (1925). Eine axiomatisierung der mengenlehre. *Journal für die reine und angewandte Mathematik* **154**, 219–240.
Vopěnka, P., & Hájek, P. (1972). *The theory of semisets*. Amsterdam: North-Holland.
Vopěnka, P. (1966). A new proof of Gödel's results on non-provability of consistency. *Bulletin de l'Académie Polonaise des Sciences* **14**(3), 111.
Wang, H. (1953). Review: Note on arithmetic models for consistent formulae of the predicate calculus by G. Kreisel. *Journal of Symbolic Logic* **18**(2), 180–181.
Wang, H. (1955). Undecidable sentences generated by semantic paradoxes. *Journal of Symbolic Logic* **20**(1), 31–43.
Wang, H. (1963). *A survey of mathematical logic*. Amsterdam: North Holland.
Wang, H. (1981). *Popular lectures on mathematical logic*. Mineola: Dover.
Weyl, H. (1918). *Das Kontinuum. Kritische Untersuchungen über die Grundlagen der Analysis*. Leipzig: Verlag von Veit & Comp.
Weyl, H. (1919). Der circulus vitiosus in der heutigen Begründung der Analysis. *Jahresbericht der Deutschen Mathematiker-Vereinigung* **28**, 85–102.
Zach, R. (1999). Completeness before Post: Bernays, Hilbert, and the development of propositional logic. *Bulletin of Symbolic Logic* **5**(03), 331–366.