THE BULLETIN OF SYMBOLIC LOGIC
Volume 27, Number 2, June 2021

CURRENT RESEARCH ON GÖDEL'S INCOMPLETENESS THEOREMS

YONG CHENG

Abstract. We give a survey of current research on Gödel's incompleteness theorems from the following three aspects: classifications of different proofs of Gödel's incompleteness theorems, the limit of the applicability of Gödel's first incompleteness theorem, and the limit of the applicability of Gödel's second incompleteness theorem.

§1. Introduction. Gödel's first and second incompleteness theorems are some of the most important and profound results in the foundations of mathematics and have had wide influence on the development of logic, philosophy, mathematics, computer science, as well as other fields. Intuitively speaking, Gödel's incompleteness theorems express that any rich enough logical system cannot prove its own *consistency*, i.e., that no contradiction like $0 = 1$ can be derived within this system.

Gödel [43] proves his first incompleteness theorem (G1) for a certain formal system **P** related to Russell–Whitehead's *Principia Mathematica* based on the simple theory of types over the natural number series and the Dedekind–Peano axioms (see [8, p. 3]). Gödel announces the second incompleteness theorem (G2) in an abstract published in October 1930: no consistency proof of systems such as Principia, Zermelo–Fraenkel set theory, or the systems investigated by Ackermann and von Neumann is possible by methods which can be formulated in these systems (see [147, p. 431]).

Gödel comments in a footnote of [43] that G2 is corollary of G1 (and in fact a formalized version of G1): if T is consistent, then the consistency of T is not provable in T where the consistency of T is formulated as the arithmetic formula which says that there exists an unprovable sentence in T. Gödel [43] sketches a proof of G2 and promises to provide full details in a subsequent publication. This promise is not fulfilled, and a detailed proof of G2 for first-order arithmetic only appears in a monograph by Hilbert and Bernays [59]. Abstract logic-free formulations of Gödel's incompleteness theorems have been given by Kleene [76] ("symmetric form"), Smullyan [121] ("representation systems"), and others. The following is a modern reformulation of Gödel's incompleteness theorems.

Received October 23, 2019.
2020 *Mathematics Subject Classification*. 03F40, 03F30, 03-02.
Key words and phrases. Gödel's first incompleteness theorem, Gödel's second incompleteness theorem, concrete incompleteness, interpretation, intensionality.

© 2021, Association for Symbolic Logic
1079-8986/21/2702-0001
DOI :10.1017/bsl.2020.44

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

THEOREM 1.1 (Gödel [43]). *Let T be a recursively axiomatized extension of* **PA**.

G1 *If T is ω-consistent, then T is incomplete.*
G2 *If T is consistent, then the consistency of T is not provable in T.*

Gödel's incompleteness theorems G1 and G2 are of a rather different nature and scope. In this paper, we will discuss different versions of G1 and G2, from incompleteness for extensions of **PA** to incompleteness for systems weaker than **PA** w.r.t. interpretation. We will freely use G1 and G2 to refer to both Gödel's first and second incompleteness theorems, and their different versions. The meaning of G1 and G2 will be clear from the context in which we refer to them.

Gödel's incompleteness theorems exhibit certain weaknesses and limitations of a given formal system. For Gödel, his incompleteness theorems indicate the creative power of human reason. In Emil Post's celebrated words: mathematical proof is an essentially creative activity (see [97, p. 339]). The impact of Gödel's incompleteness theorems is not confined to the community of mathematicians and logicians; popular accounts are well-known within the general scientific community and beyond. Gödel's incompleteness theorems raise a number of philosophical questions concerning the nature of logic and mathematics as well as mind and machine. For the impact of Gödel's incompleteness theorems, Feferman said:

> their relevance to mathematical logic (and its offspring in the theory of computation) is paramount; further, their philosophical relevance is significant, but in just what way is far from settled; and finally, their mathematical relevance outside of logic is very much unsubstantiated but is the object of ongoing, tantalizing efforts (see [34, p. 434]).

From the literature, there are some good textbooks and survey papers on Gödel's incompleteness theorems. For textbooks, we refer to [14, 29, 37, 40, 53, 91, 97, 118, 120, 121]. For survey papers, we refer to [8, 12, 16, 23, 79, 119, 133]. In the last 20 years, there have been a lot of advances in the study of incompleteness. We felt that a comprehensive survey paper for the current state-of-art of this research field is missing from the literature. The motivation of this paper is four-fold:

- Give the reader an overview of the current state-of-art of research on incompleteness.
- Classify these new advances on incompleteness under some important themes.
- Propose some new questions not covered in the literature.
- Set the direction for the future research of incompleteness.

Due to space limitations and our personal taste, it is impossible to cover all research results from the literature related to incompleteness in this survey. Therefore, we will focus on three aspects of new advances in research on incompleteness:

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

- classifications of different proofs of Gödel's incompleteness theorems;
- the limit of the applicability of G1;
- the limit of the applicability of G2.

We think these are the most important three aspects of research on incompleteness and reflect the depth and breadth of the research on incompleteness after Gödel. In this survey, we will focus on logical and mathematical aspects of research on incompleteness.

An important and interesting topic concerning incompleteness is missing in this paper: philosophy of Gödel's incompleteness theorems. For us, the widely discussed and most important philosophical questions about Gödel's incompleteness theorems are: the relationship between G1 and the mechanism thesis, the status of Gödel's disjunctive thesis, and the intensionality problem of G2. We leave a survey of philosophical discussions of Gödel's incompleteness theorems for a future philosophy paper.

This paper is structured as follows. In Section 1, we introduce the motivation, the main content, and the structure of this paper. In Section 2, we list the preliminary notions and definitions used in this paper. In Section 3, we examine different proofs of Gödel's incompleteness theorems and classify these proofs based on nine criteria. In Section 4, we examine the limit of the applicability of G1 both for extensions of **PA**, and for theories weaker than **PA** w.r.t. interpretation. In Section 5, we examine the limit of the applicability of G2, and discuss sources of indeterminacy in the formulation of the consistency statement.

§2. Preliminaries.

2.1. ***Definitions and notations.*** We list the definitions and notations required below. These are standard and used throughout the literature.

DEFINITION 2.1 (Basic notions).

- A *language* consists of an arbitrary number of relation and function symbols of arbitrary finite arity.[1] For a given theory T, we use $L(T)$ to denote the language of T, and often equate $L(T)$ with the list of non-logical symbols of the language.
- For a formula ϕ in $L(T)$, '$T \vdash \phi$' denotes that ϕ is provable in T: i.e., there is a finite sequence of formulas $\langle \phi_0, \ldots, \phi_n \rangle$ such that $\phi_n = \phi$, and for any $0 \leq i \leq n$, either ϕ_i is an axiom of T, or ϕ_i follows from some ϕ_j $(j < i)$ by using one inference rule.
- A theory T is *consistent* if no contradiction is provable in T.
- We say a sentence ϕ is *independent* of T if $T \nvdash \phi$ and $T \nvdash \neg\phi$.
- A theory T is *incomplete* if there is a sentence ϕ in $L(T)$ which is independent of T; otherwise, T is *complete* (i.e., for any sentence ϕ in $L(T)$, either $T \vdash \phi$ or $T \vdash \neg\phi$).

[1] We may view nullary functions as constants, and nullary relations as propositional variables.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

In this paper, we focus on first-order theories based on a countable language, and always assume the *arithmetization* of the base theory with a recursive set of non-logical symbols. For the technical details of arithmetization, we refer to [18, 97]. Arithmetization means that any formula or finite sequence of formulas can be coded by a natural number, called the *Gödel number*. This representation of syntax was pioneered by Gödel.

DEFINITION 2.2 (Basic notions following arithmetization).

- We say a set of sentences Σ is *recursive* if the set of Gödel numbers of sentences in Σ is recursive.
- A theory T is *decidable* if the set of sentences provable in T is recursive; otherwise it is *undecidable*.
- A theory T is *recursively axiomatizable* if it has a recursive set of axioms (i.e., the set of Gödel numbers of axioms of T is recursive).
- A theory T is *finitely axiomatizable* if it has a finite set of axioms.
- A theory T is *locally finitely satisfiable* if every finitely axiomatized subtheory of T has a finite model.
- A theory T is *recursively enumerable* (r.e.) if it has a recursively enumerable set of axioms.
- A theory T is *essentially undecidable* if any recursively axiomatizable consistent extension of T in the same language is undecidable.
- A theory T is *essentially incomplete* if any recursively axiomatizable consistent extension of T in the same language is incomplete.[2]
- A theory T is *minimal essentially undecidable* if T is essentially undecidable, and if deleting any axiom of T, the remaining theory is no longer essentially undecidable.

DEFINITION 2.3 (Basic notations).

- We denote by $\overline{n}$ the numeral representing $n \in \omega$ in $L(\mathbf{PA})$.
- We denote by $\ulcorner \phi \urcorner$ the numeral representing the Gödel number of ϕ.
- We denote by $\ulcorner \phi(\dot{x}) \urcorner$ the numeral representing the Gödel number of the sentence obtained by replacing x with the value of x.[3]

DEFINITION 2.4 (Representations, translations, and interpretations).

- A n-ary relation $R(x_1, \ldots, x_n)$ on ω^n is *representable* in T if there is a formula $\phi(x_1, \ldots, x_n)$ such that $T \vdash \phi(\overline{m_1}, \ldots, \overline{m_n})$ when $R(m_1, \ldots, m_n)$ holds, and $T \vdash \neg\phi(\overline{m_1}, \ldots, \overline{m_n})$ when $R(m_1, \ldots, m_n)$ does not hold.
- We say that a total function $f(x_1, \ldots, x_n)$ on ω^n is *representable in T* if there is a formula $\varphi(x_1, \ldots, x_n, y)$ such that $T \vdash \forall y(\varphi(\overline{a_1}, \ldots, \overline{a_n}, y) \leftrightarrow y = \overline{m})$ whenever $a_1, \ldots, a_n, m \in \omega$ are such that $f(a_1, \ldots, a_n) = m$.
- Let T be a theory in a language $L(T)$, and S a theory in a language $L(S)$. In its simplest form, a *translation* I of language $L(T)$ into language $L(S)$ is specified by the following:

[2] The theory of completeness/incompleteness is closely related to the theory of decidability/undecidability (see [125]).

[3] Note that the variable x is free in the formula $\ulcorner \phi(\dot{x}) \urcorner$ but not in $\ulcorner \phi(x) \urcorner$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

  - an $L(S)$-formula $\delta_I(x)$ denoting the domain of I;
  - for each relation symbol R of $L(T)$, as well as the equality relation $=$, an $L(S)$-formula R_I of the same arity;
  - for each function symbol F of $L(T)$ of arity k, an $L(S)$-formula F_I of arity $k+1$.
- If ϕ is an $L(T)$-formula, its I-translation ϕ^I is an $L(S)$-formula constructed as follows: we rewrite the formula in an equivalent way so that function symbols only occur in atomic subformulas of the form $F(\overline{x}) = y$, where x_i, y are variables; then we replace each such atomic formula with $F_I(\overline{x}, y)$, we replace each atomic formula of the form $R(\overline{x})$ with $R_I(\overline{x})$, and we restrict all quantifiers and free variables to objects satisfying δ_I. We take care to rename bound variables to avoid variable capture during the process.
- A translation I of $L(T)$ into $L(S)$ is an *interpretation* of T in S if S proves the following:
  - for each function symbol F of $L(T)$ of arity k, the formula expressing that F_I is total on δ_I:

$$\forall x_0, \ldots, \forall x_{k-1}(\delta_I(x_0) \wedge \cdots \wedge \delta_I(x_{k-1}) \rightarrow \exists y(\delta_I(y) \\ \wedge F_I(x_0, \ldots, x_{k-1}, y)));$$

  - the I-translations of all axioms of T, and axioms of equality.

The simplified picture of translations and interpretations above actually describes only *one-dimensional*, *parameter-free*, and *one-piece* translations. For precise definitions of a *multi-dimensional interpretation*, an *interpretation with parameters*, and a *piece-wise interpretation*, we refer to [131–133] for more details.

The notion of interpretation provides us with a method for comparing different theories in different languages, as follows.

DEFINITION 2.5 (Interpretations II).

- A theory T is *interpretable* in a theory S if there exists an interpretation of T in S. If T is interpretable in S, then all sentences provable (refutable) in T are mapped, by the interpretation function, to sentences provable (refutable) in S.
- We say that a theory U *weakly interprets* a theory V (or V is *weakly interpretable* in U) if V is interpretable in some consistent extension of U in the same language (or equivalently, for some interpretation τ, the theory $U + V^\tau$ is consistent).
- Given theories S and T, let '$S \trianglelefteq T$' denote that S is interpretable in T (or T interprets S); let '$S \triangleleft T$' denote that T interprets S but S does not interpret T; we say S and T are *mutually interpretable* if $S \trianglelefteq T$ and $T \trianglelefteq S$.

Interpretability provides us with one measure of comparing strength of different theories. If theories S and T are mutually interpretable, then T and S are equally strong w.r.t. interpretation. In this paper, whenever we say that theory S is weaker than theory T w.r.t. interpretation, this means that $S \triangleleft T$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

A general method for establishing the undecidability of theories is developed in [125]. The following theorem provides us with two methods for proving the essentially undecidability of a theory respectively via interpretation and representability.

THEOREM 2.6 ([125, Theorem 7 and Corollary 2]).

- *Let T_1 and T_2 be two consistent theories such that T_2 is interpretable in T_1. If T_2 is essentially undecidable, then T_1 is essentially undecidable.*
- *If all recursive functions are representable in a consistent theory T, then T is essentially undecidable.*

We shall also need some basic notions from recursion theory, as follows.

DEFINITION 2.7 (Basic recursion theory).

- Let $\phi_0, \phi_1, \ldots$ be a list of all unary computable (partial recursive) functions such that $\phi_i(j)$, if it exists, can be computed from i and j.
- A recursively enumerable set (r.e. for short) is the domain of ϕ_i for some $i \in \omega$, which is denoted by W_i.
- The notation $\phi_i(j)\uparrow$ means that the function ϕ_i is not defined at j, or $j \notin W_i$; and $\phi_i(j)\downarrow$ means that ϕ_i is defined at j, or $j \in W_i$.

Provability logic provides us with an important tool to study the metamathematics of arithmetic and incompleteness. A good reference on the basics of provability logic is [14].

DEFINITION 2.8 (Modal logic).

- The modal system **K** consisting of the following axiom schemes:
  - All tautologies;
  - $\Box(A \to B) \to (\Box A \to \Box B)$;

  as well as two inference rules:
  - if $\vdash A$ and $\vdash A \to B$, then $\vdash B$;
  - if $\vdash A$, then $\vdash \Box A$.
- We denote by **GL** the modal system consisting of all axioms of **K**, all instances of the scheme $\Box(\Box A \to A) \to \Box A$, and the same inference rules with **K**.
- We denote by **GLS** the modal system consisting of all theorems of **GL**, and all instances of the scheme $\Box A \to A$. However, **GLS** has only one inference rule: Modus Ponens.

2.2. *Logical systems.* In this section, we introduce some well-known theories weaker than **PA** w.r.t. interpretation from the literature. In Section 4, we will show that these theories are essentially incomplete.

Robinson Arithmetic **Q** is introduced in [125] by Tarski, Mostowski, and Robinson as a base axiomatic theory for investigating incompleteness and undecidability.

DEFINITION 2.9 (Robinson Arithmetic **Q**). Robinson Arithmetic **Q** is defined in the language $\{\mathbf{0}, \mathbf{S}, +, \times\}$ with the following axioms:

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

$\mathbf{Q}_1$: $\forall x \forall y(\mathbf{S}x = \mathbf{S}y \to x = y)$;
$\mathbf{Q}_2$: $\forall x(\mathbf{S}x \neq \mathbf{0})$;
$\mathbf{Q}_3$: $\forall x(x \neq \mathbf{0} \to \exists y(x = \mathbf{S}y))$;
$\mathbf{Q}_4$: $\forall x \forall y(x + \mathbf{0} = x)$;
$\mathbf{Q}_5$: $\forall x \forall y(x + \mathbf{S}y = \mathbf{S}(x + y))$;
$\mathbf{Q}_6$: $\forall x(x \times \mathbf{0} = \mathbf{0})$;
$\mathbf{Q}_7$: $\forall x \forall y(x \times \mathbf{S}y = x \times y + x)$.

Robinson Arithmetic **Q** is very weak: it cannot even prove that addition is associative.

DEFINITION 2.10 (Peano Arithmetic **PA**). The theory **PA** consists of the axioms $\mathbf{Q}_1$ and $\mathbf{Q}_2$, $\mathbf{Q}_4$–$\mathbf{Q}_7$ in Definition 2.9, and the following axiom scheme of induction:

$$(\phi(\mathbf{0}) \wedge \forall x(\phi(x) \to \phi(\mathbf{S}x))) \to \forall x \phi(x), \qquad \text{(Induction)}$$

where ϕ is a formula with at least one free variable x. Let $\mathfrak{N} = \langle \mathbb{N}, +, \times \rangle$ denote the standard model of arithmetic.

We now introduce a well-known hierarchy of $L(\mathbf{PA})$-formulas called the *arithmetical hierarchy* (see [53, 97]).

DEFINITION 2.11 (Arithmetical hierarchy).

- *Bounded formulas* (Σ^0_0, or Π^0_0, or Δ^0_0 formulas) are built from atomic formulas using only propositional connectives and bounded quantifiers (in the form $\forall x \leq y$ or $\exists x \leq y$).
- A formula is Σ^0_{n+1} if it has the form $\exists x \phi$ where ϕ is Π^0_n.
- A formula is Π^0_{n+1} if it has the form $\forall x \phi$ where ϕ is Σ^0_n. Thus, a Σ^0_n-formula has a block of n alternating quantifiers, the first one being existential, and this block is followed by a bounded formula. Similarly for Π^0_n-formulas.
- A formula is Δ^0_n if it is equivalent to both a Σ^0_n formula and a Π^0_n formula.

We can now formally introduce the notion of consistency in its various guises, as well as the various fragments of Peano Arithmetic **PA**.

DEFINITION 2.12 (Formal consistency and systems).

- A theory T is said to be ω *-consistent* if there is no formula $\varphi(x)$ such that $T \vdash \exists x \varphi(x)$, and for any $n \in \omega$, $T \vdash \neg\varphi(\overline{n})$.
- A theory T is 1 *-consistent* if there is no such Δ^0_1 formula $\varphi(x)$.
- We say a theory T is Σ^0_1 *-sound* if for any Σ^0_1 sentences ϕ, if $T \vdash \phi$, then $\mathfrak{N} \models \phi$.
- The collection axiom for Σ^0_{n+1} formulas is the following principle: $(\forall x < u)(\exists y)\varphi(x,y) \to (\exists v)(\forall x < u)(\exists y < v)\varphi(x,y)$ where $\varphi(x,y)$ is a Σ^0_{n+1} formula possibly containing parameters distinct from u, v.
- The theory $I\Sigma_n$ is **Q** plus induction for Σ^0_n formulas, and $B\Sigma_{n+1}$ is $I\Sigma_0$ plus collection for Σ^0_{n+1} formulas.
- The theory $I\Delta_0$ is **Q** plus induction for Δ^0_0 formulas.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

- The theory **PA** is the union of all $I\Sigma_n$.

It is well-known that the following form a strictly increasing hierarchy:

$$I\Sigma_0, B\Sigma_1, I\Sigma_1, B\Sigma_2, \ldots, I\Sigma_n, B\Sigma_{n+1}, \ldots, \mathbf{PA}.$$

Moreover, there are *weak* fragments of **PA** that play an important role in computer science, namely in *complexity theory* [17, 18]. These systems are based on the following concept.

By [35, Proposition 2, p. 299], there is a bounded formula $\mathsf{Exp(x, y, z)}$ such that $I\Sigma_0$ proves that $\mathsf{Exp(x, 0, z)} \leftrightarrow z = 1$, and $\mathsf{Exp(x, Sy, z)} \leftrightarrow \exists t(\mathsf{Exp(x, y, t)} \wedge z = t \cdot x)$. However, $I\Sigma_0$ cannot prove the totality of $\mathsf{Exp(x, y, z)}$.

Definition 2.13 (Sub-exponential functions).

- Let **exp** denote the statement postulating the totality of the exponential function $\forall x \forall y \exists z \mathsf{Exp(x, y, z)}$.
- Elementary Arithmetic (**EA**) is $I\Delta_0 + \mathbf{exp}$.
- Define $\omega_1(x) = x^{|x|}$, and $\omega_{n+1}(x) = 2^{\omega_n(|x|)}$ where $|x|$ is the length of the binary expression of x.
- Let $\Omega_n \equiv (\forall x)(\exists y)(\omega_n(x) = y)$ express that $\omega_n(x)$ is total.

Theorem 2.14 ([35, 51]).

- *The theory $I\Sigma_0 + \Omega_n$ is interpretable in* **Q** *for any $n \geq 1$ (see [35, Theorem 3, p. 304]).*
- *The theory $I\Sigma_0 + \mathbf{exp}$ is not interpretable in* **Q**.[4]
- *The theory $I\Sigma_1$ is not interpretable in $I\Sigma_0 + \mathbf{exp}$ (see [51, Theorem 1.1, p. 186]).*
- *The theory $I\Sigma_{n+1}$ is not interpretable in $B\Sigma_{n+1}$ (see [51, Theorem 1.2, p. 186]).*
- *The theory $B\Sigma_1 + exp$ is interpretable in $I\Sigma_0 + \mathbf{exp}$ (see [51, Theorem 2.4, p. 188]).*
- *The theory $B\Sigma_1 + \Omega_n$ is interpretable in $I\Sigma_0 + \Omega_n$ for each $n \geq 1$ (see [51, Theorem 2.5, p. 189]).*
- *The theory $B\Sigma_{n+1}$ is interpretable in $I\Sigma_n$ for each $n \geq 0$ (see [51, Theorem 2.6, p. 189]).*

The theory $\mathbf{PA}^-$ is the theory of commutative, discretely ordered semirings with a minimal element plus the subtraction axiom. The theory $\mathbf{PA}^-$ has the following axioms, where the language $L(\mathbf{PA}^-)$ is $L(\mathbf{PA}) \cup \{\leq\}$:

Definition 2.15 (The system $\mathbf{PA}^-$).

- $x + 0 = x$;
- $x + y = y + x$;
- $(x + y) + z = x + (y + z)$;
- $x \times 1 = x$;
- $x \times y = y \times x$;
- $(x \times y) \times z = x \times (y \times z)$;

[4]See [35, Theorem 6, p. 313]. Solovay proves that $I\Sigma_0 + \neg\mathbf{exp}$ is interpretable in **Q** (see [35, Theorem 7, p. 314]).

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

- $x \times (y+z) = x \times y + x \times z$;
- $x \leq y \vee y \leq x$;
- $(x \leq y \wedge y \leq z) \rightarrow x \leq z$;
- $x+1 \not\leq x$;
- $x \leq y \rightarrow (x = y \vee x+1 \leq y)$;
- $x \leq y \rightarrow x+z \leq y+z$;
- $x \leq y \rightarrow x \times z \leq y \times z$;
- $x \leq y \rightarrow \exists z(x+z = y)$.

The theory $\mathbf{Q}^+$ is the extension of $\mathbf{Q}$ in the language $L(\mathbf{Q}^+) = L(\mathbf{Q}) \cup \{\leq\}$ with the following extra axioms:

Definition 2.16 (The system $\mathbf{Q}^+$). The system $\mathbf{Q}^+$ is $\mathbf{Q}$ plus:

$\mathbf{Q}_8$: $(x+y)+z = x+(y+z)$;
$\mathbf{Q}_9$: $x \times (y+z) = x \times y + x \times z$;
$\mathbf{Q}_{10}$: $(x \times y) \times z = x \times (y \times z)$;
$\mathbf{Q}_{11}$: $x+y = y+x$;
$\mathbf{Q}_{12}$: $x \times y = y \times x$;
$\mathbf{Q}_{13}$: $x \leq y \leftrightarrow \exists z(x+z = y)$.

Andrzej Grzegorczyk considers a theory $\mathbf{Q}^-$ in which addition and multiplication satisfy natural reformulations of the axioms of $\mathbf{Q}$ but are possibly *non-total* functions. More exactly, the language of $\mathbf{Q}^-$ is $\{\mathbf{0}, \mathbf{S}, A, M\}$ where A and M are ternary relations.

Definition 2.17 (The system $\mathbf{Q}^-$). The axioms of $\mathbf{Q}^-$ are the axioms $\mathbf{Q}_1$–$\mathbf{Q}_3$ of $\mathbf{Q}$ plus the following six axioms about A and M:

A: $\forall x \forall y \forall z_1 \forall z_2 (A(x,y,z_1) \wedge A(x,y,z_2) \rightarrow z_1 = z_2)$;
M: $\forall x \forall y \forall z_1 \forall z_2 (M(x,y,z_1) \wedge M(x,y,z_2) \rightarrow z_1 = z_2)$;
G4: $\forall x\, A(x,0,x)$;
G5: $\forall x \forall y \forall z (\exists u (A(x,y,u) \wedge z = S(u)) \rightarrow A(x,S(y),z))$;
G6: $\forall x\, M(x,0,0)$;
G7: $\forall x \forall y \forall z (\exists u (M(x,y,u) \wedge A(u,x,z)) \rightarrow M(x,S(y),z))$.

Samuel R. Buss [17] introduces $\mathbf{S}^1_2$, a finitely axiomatizable theory, to study polynomial time computability. The theory $\mathbf{S}^1_2$ provides what is needed for formalizing the proof of G2 in a natural and effortless way: this process is actually easier in Buss' theory than in full **PA**, since the restrictions present in $\mathbf{S}^1_2$ prevent one from making wrong turns and inefficient choices (see [132]).

Next, we introduce *adjunctive set theory* **AS** which has a language with only one binary relation symbol '$\in$'.

Definition 2.18 (Adjunctive set theory **AS**, [98]). The axioms of **AS** consist of the following:

AS1: $\exists x \forall y (y \notin x)$.
AS2: $\forall x \forall y \exists z \forall u (u \in z \leftrightarrow (u = x \vee u = y))$.

We now consider the theory **R** introduced by A. Tarski, A. Mostowski, and R. Robinson in [125], and some variants of it.

Definition 2.19 (The theory **R**). Let **R** be the theory consisting of schemes Ax1– Ax5 with $L(\mathbf{R}) = \{\overline{0}, \ldots, \overline{n}, \ldots, +, \times, \leq\}$ where $m, n \in \omega$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Ax1: $\overline{m}+\overline{n}=\overline{m+n}$;
Ax2: $\overline{m}\times\overline{n}=\overline{m\times n}$;
Ax3: $\overline{m}\neq\overline{n}$ if $m\neq n$;
Ax4: $\forall x(x\leq\overline{n}\rightarrow x=\overline{0}\vee\cdots\vee x=\overline{n})$;
Ax5: $\forall x(x\leq\overline{n}\vee\overline{n}\leq x)$.

As it happens, the system **R** contains all key properties of arithmetic for the proof of G1. Unlike **Q**, the theory **R** is not finitely axiomatizable.

Definition 2.20 (Variations of **R**).

- Let $\mathbf{R}_0$ be **R** without **Ax5**.
- Let $\mathbf{R}_1$ be the system consisting of schemes **Ax1**, **Ax2**, **Ax3**, and **Ax4′** where the latter is as follows:
  Ax4′ $\forall x(x\leq\overline{n}\leftrightarrow x=\overline{0}\vee\cdots\vee x=\overline{n})$.
- Let $\mathbf{R}_2$ be the system consisting of schemes **Ax2**, **Ax3**, and **Ax4′**.

The 'concatenation' theory **TC** has the language $\{\frown,\alpha,\beta\}$ with a binary function symbol and two constants.

Definition 2.21 (The system **TC**).

TC1: $\forall x\forall y\forall z(x\frown(y\frown z)=(x\frown y)\frown z)$;
TC2: $\forall x\forall y\forall u\forall v(x\frown y=u\frown v\rightarrow((x=u\wedge y=v)\vee\exists w((u=x\frown w\wedge w\frown v=y)\vee(x=u\frown w\wedge w\frown y=v))))$;
TC3: $\forall x\forall y(\alpha\neq x\frown y)$;
TC4: $\forall x\forall y(\beta\neq x\frown y)$;
TC5: $\alpha\neq\beta$.

Primitive recursive arithmetic (**PRA**) is a quantifier-free formalization of the natural numbers, and the language of **PRA** can express arithmetic statements involving natural numbers and any primitive recursive function. Weak Konig's Lemma ($\mathbf{WKL_0}$) states that every infinite binary tree has an infinite branch. We refer to [53, 117] for the definitions of **PRA** and $\mathbf{WKL_0}$. In a nutshell, the former system allows us to perform 'iteration of functions $f:\mathbb{N}\rightarrow\mathbb{N}$', while the latter expresses a basic compactness argument for Cantor space.

Theorem 2.22 (Friedman's conservation theorem [74, Theorem 2.1]). *For any Π^0_2 sentence ϕ in $L(\mathbf{PA})$, if $\mathbf{WKL_0}\vdash\phi$, then $\mathbf{PRA}\vdash\phi$.*

Finally, diagnolisation, in one form or other, forms the basis for the proof of G2. The following lemma is crucial in this regard.

Lemma 2.23 (The Diagnolisation Lemma). *Let T be a consistent r.e. extension of **Q**. For any formula $\phi(x)$ with exactly one free variable, there exists a sentence θ such that $T\vdash\theta\leftrightarrow\phi(\ulcorner\theta\urcorner)$.*

Lemma 2.23 is the simplest and most often used version of the Diagnolisation Lemma. For a generalized version of the Diagnolisation Lemma, we refer to [14]. In this paper, we use the term "Diagnolisation Lemma" to refer to Lemma 2.23 and some variants of the generalized version.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

§3. Proofs of Gödel's incompleteness theorems.

3.1. ***Introduction.*** In this section, we discuss different proofs of Gödel's incompleteness theorems from the literature, and propose nine criteria for classifying them.

First of all, there are no requirements on the independent sentence in G1. In particular, such a sentence need not have any mathematical meaning. This is often the case when meta-mathematical (proof-theoretic or recursion-theoretic or model-theoretic) methods are used to construct the independent sentence. In Sections 3.2–3.4, we will discuss proofs of Gödel's incompleteness theorems *via pure logic*. In Section 3.5, we will give an overview of the "concrete incompleteness" research program which seeks to identify natural independent sentences *with real mathematical meaning*.

Secondly, we say that a proof of G1 is *constructive* if it explicitly constructs the independent sentence from the base theory by algorithmic means. A non-constructive proof of G1 only proves the mere existence of the independent sentence and does not show its existence algorithmically. We say that a proof of G1 for theory T has *the Rosser property* if the proof only assumes that T is consistent instead of assuming that T is ω-consistent or 1-consistent or Σ^0_1-sound; all these notions are introduced in Section 2.2.

After Gödel, many different proofs of Gödel's incompleteness theorems have been found. These proofs can be classified using the following criteria:

- proof-theoretic proof;
- recursion-theoretic proof;
- model-theoretic proof;
- proof via arithmetization;
- proof via the Diagnolisation Lemma;
- proof based on "logical paradox";
- constructive proof;
- proof having the Rosser property;
- the independent sentence has natural and real mathematical content.[5]

However, these aspects are not exclusive: a proof of G1 or G2 may satisfy several of the above criteria.

Thirdly, there are two kinds of proofs of Gödel's incompleteness theorems via pure logic: one based on logical paradox and one not based on logical paradox. In Section 3.2, we first provide an overview of the modern reformulation of proofs of Gödel's incompleteness theorems. We discuss proofs of Gödel's incompleteness theorems not based on logical paradox in Section 3.3. We discuss proofs of Gödel's incompleteness theorems based on logical paradox in Section 3.4.

[5]I.e., Gödel's sentence is a pure logical construction (via the arithmetization of syntax and provability predicate) and has no relevance with classic mathematics (without any combinatorial or number-theoretic content). On the contrary, Paris–Harrington Principle is an independent arithmetic sentence from classic mathematics with combinatorial content.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

3.2. *Overview and modern formulation.* In a nutshell, the three main ideas in the (modern/standard) proofs of G1 and G2 are *arithmetization*, *representability*, and *self-reference*, as discussed in detail in Section 3.2.1. Interesting properties of G1 and G2 are discussed in Sections 3.2.2 and 3.2.4, while the formalized notions of 'proof' and 'truth' are discussed in Section 3.2.3. Finally, we formulate a blanket caveat for the rest of this section:

> *Unless stated otherwise, we will always assume that T is a recursively axiomatizable consistent extension of* **Q**.

Other sections shall contain similar caveats and we sometimes stress these.

3.2.1. Three steps towards G1 *and* G2. Intuitively speaking, Gödel's incompleteness theorems can be proved based on the following key ingredients.

- **Arithmetization**: since G1 and G2 are theorems about properties of the syntax of logic, we need to somehow represent the latter, which is done via a coding scheme called *arithmetization*.
- **Representations**: the notion of 'proof' and related concepts in G1 and G2 are then expressed ('represented') via arithmetization.
- **Self-reference**: given a representation of 'proof' and related concepts, one can write down formal statements that intuitively express 'self-referential' things like 'this sentence does not have a proof'.

As we will see, the intuitively speaking 'self-referential' statements are the key to proving G1 and G2. We now discuss these three notions in detail.

First of all, **arithmetization** has the following intuitive content: it establishes a one-to-one correspondence between expressions of $L(T)$ and natural numbers. Thus, we can translate metamathematical statements about the formal theory T into statements about natural numbers. Furthermore, fundamental metamathematical relations can be translated in this way into certain recursive relations, hence into relations representable in T. Consequently, one can speak about a formal system of arithmetic, and about its properties as a theory in the system itself (see [97])! This is the essence of Gödel's idea of arithmetization, which was revolutionary at a time when computer hardware and software did not exist yet.

Secondly, in light of the previous, we can define certain relations on natural numbers that express or **represent** crucial metamathematical concepts related to the formal system T, like 'proof' and 'consistency'. For example, modulo plenty of technical details, we can readily define a binary relation on ω^2 expressing what it means to prove a formula in T, namely as follows:

> $Proof_T(m,n)$ if and only if n is the Gödel number of a proof in T of the formula with Gödel number m.

Moreover, we can show that the relation $Proof_T(m,n)$ is recursive. In addition, Gödel proves that every recursive relation is representable in **PA**.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Next, let $\mathbf{Proof}_T(x,y)$ be the formula which represents $Proof_T(m,n)$ in **PA**.[6] From the formula $\mathbf{Proof}_T(x,y)$, we can define the 'provability' predicate $\mathbf{Prov}_T(x)$ as $\exists y \mathbf{Proof}_T(x,y)$. The provability predicate $\mathbf{Prov}_T(x)$ satisfies the following conditions which show that formal and intuitive provability have the same properties.

(1) If $T \vdash \varphi$, then $T \vdash \mathbf{Prov}_T(\ulcorner \varphi \urcorner)$;
(2) $T \vdash \mathbf{Prov}_T(\ulcorner \varphi \to \psi \urcorner) \to (\mathbf{Prov}_T(\ulcorner \varphi \urcorner) \to \mathbf{Prov}_T(\ulcorner \psi \urcorner))$;
(3) $T \vdash \mathbf{Prov}_T(\ulcorner \varphi \urcorner) \to \mathbf{Prov}_T(\ulcorner \mathbf{Prov}_T(\ulcorner \varphi \urcorner) \urcorner)$.

For the proof of G1, Gödel defines the *Gödel sentence* **G** which asserts its own unprovability in T via a **self-reference** construction. Gödel shows that if T is consistent, then $T \nvdash \mathbf{G}$, and if T is ω-consistent, then $T \nvdash \neg\mathbf{G}$. One way of obtaining such a Gödel sentence is the *Diagnolisation Lemma* which intuitively speaking implies that the predicate $\neg\mathbf{Prov}_T(x)$ has a *fixed point*, i.e., there is a sentence θ in $L(T)$ such that

$$T \vdash \theta \leftrightarrow \neg\mathbf{Prov}_T(\ulcorner \theta \urcorner).$$

Clearly, $T \nvdash \theta$ while θ intuitively expresses its own unprovability, i.e., the aforementioned **self-referential** nature.

For the proof of G2, we first define the arithmetic sentence $\mathbf{Con}(T)$ in $L(T)$ as $\neg\mathbf{Prov}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$ which says that for all x, x is not a code of a proof of a contradiction in T. Gödel's second incompleteness theorem (G2) states that if T is consistent, then the arithmetical formula $\mathbf{Con}(T)$, which expresses the consistency of T, is not provable in T. In Section 5.3, we will discuss some other ways of expressing the consistency of T.

Finally, from the above conditions (1)–(3), one can show that $T \vdash \mathbf{Con}(T) \leftrightarrow \mathbf{G}$. Thus, G2 holds: if T is consistent, then $T \nvdash \mathbf{Con}(T)$. For more details on these proofs of G1 and G2, we refer to Chapter 2 in [97].

3.2.2. Properties of G1. In this section, we discuss some (sometimes subtle) comments on G1.

First of all, Gödel's proof of G1 is constructive as follows: given a consistent r.e. extension T of **PA**, the proof constructs, in an algorithmic way, a true arithmetic sentence which is unprovable in T. In fact, one can effectively find a true Π^0_1 sentence G_T of arithmetic such that G_T is independent of T. Gödel calls this the "incompletability or inexhaustability of mathematics".

Secondly, for Gödel's proof of G1, only assuming that T is consistent does not suffice to show that Gödel sentence is independent of T. In fact, the optimal condition to show that Gödel sentence is independent of T is: $T + \mathbf{Con}(T)$ is consistent (see Theorems 35 and 36 in [62]).[7]

[6] Via arithmetization and representability, one can speak about the property of T in **PA** itself!

[7] This optimal condition is much weaker than ω-consistency.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Thirdly, in summary, Gödel's proof of G1 has the following properties:

- uses proof-theoretic method with arithmetization;
- does not directly use the Diagnolisation Lemma;
- the proof formalizes the liar paradox;
- the proof is constructive;
- the proof does not have the Rosser property;
- Gödel's sentence has no real mathematical content.

All these characteristics of Gödel's proof of G1 are not necessary conditions for proving G1. For example, G1 can be proved using recursion-theoretic or model-theoretic method, using the Diagnolisation Lemma, using other logical paradoxes, using non-constructive methods, only assuming that T is consistent (i.e., having the Rosser property), and can be proved without arithmetization.

Fourth, G1 does *not* tell us that any consistent theory is incomplete. In fact, there are many consistent complete first-order theories. For example, the following first-order theories are complete: the theory of dense linear orderings without endpoints (**DLO**), the theory of ordered divisible groups (**ODG**), the theory of algebraically closed fields of given characteristic ($\mathbf{ACF_p}$), and the theory of real closed fields (**RCF**). We refer to [31] for details of these theories. In fact, G1 only tells us that any consistent first-order theory containing a large enough fragment of **PA** (such as **Q**) is incomplete: there is then a true Π^0_1 sentence which is independent of the initial theory. Turing's work in [126] shows that any true Π^0_1-sentence of arithmetic is provable in some transfinite iteration of **PA**. Feferman's work in [33] extends Turing's work and shows that any true sentence of arithmetic is provable in some transfinite iteration of **PA**.

Fifth, whether a theory of arithmetic is complete depends on the language of the theory. There are respectively recursively axiomatized complete arithmetic theories in the language of $L(\mathbf{0},\mathbf{S})$, $L(\mathbf{0},\mathbf{S},<)$, and $L(\mathbf{0},\mathbf{S},<,+)$ (see Sections 3.1 and 3.2 in [29]). Containing enough information of arithmetic is essential for a consistent arithmetic theory to be incomplete. For example, Euclidean geometry is not about arithmetic but only about points, circles, and lines in general; but Euclidean geometry is complete as Tarski has proved (see [124]). If the theory contains only information about the arithmetic of addition without multiplication, then it can be complete. For example, Presburger arithmetic is a complete theory of the arithmetic of addition in the language of $L(\mathbf{0},\mathbf{S},+)$ (see [97, Theorem 3.2.2, p. 222]). Finally, containing the arithmetic of multiplication is not sufficient for a theory to be incomplete. For example, there exists a complete recursively axiomatized theory in the language of $L(\mathbf{0},\times)$ (see [97, p. 230]).

Finally, it is well-known that $Th(\mathbb{N},+,\times)$ is interpretable in $Th(\mathbb{Z},+,\times)$ and $Th(\mathbb{Q},+,\times)$.[8] Since $Th(\mathbb{N},+,\times)$ is undecidable and has a finitely axiomatizable incomplete sub-theory **Q**, by Theorem 2.6, $Th(\mathbb{Z},+,\times)$ and $Th(\mathbb{Q},+,\times)$ are undecidable, and hence not recursively axiomatizable, but

[8]The key point is: $\mathbb{N}$ is definable in $(\mathbb{Z},+,\times)$ and $(\mathbb{Q},+,\times)$. See Chapter 16 in [31].

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

they respectively have a finitely axiomatizable incomplete sub-theory of integers and rational numbers. But $Th(\mathbb{R}, +, \times)$ is decidable and recursively axiomatizable (even if not finitely axiomatizable). In fact, $Th(\mathbb{R}, +, \times) =$ **RCF** (the theory of real closed field) (see [31, p. 320–321]). Note that this fact does not contradict G1 since none of $\mathbb{N}, \mathbb{Z}$, and $\mathbb{Q}$ is definable in $(\mathbb{R}, +, \times)$.

3.2.3. Between truth and provability. In this paper, unless stated otherwise, we equate a set of sentences with the set of Gödel's numbers of these sentences. We discuss the formalized notions of 'truth' and 'proof', and how they relate to incompleteness.

DEFINITION 3.1 We define **Truth** $= \{\phi \in L(\mathbf{PA}) : \mathfrak{N} \models \phi\}$ and **Prov** $= \{\phi \in L(\mathbf{PA}) : \mathbf{PA} \vdash \phi\}$, i.e., the formalized notions of 'proof' and 'truth'.

First of all, truth and provability are the same for *purely existential statements*. Put another way, incompleteness does not arise at the level of Σ^0_1 sentences. Indeed, we have Σ^0_1-completeness for T: for any Σ^0_1 sentences ϕ, $T \vdash \phi$ if and only if $\mathfrak{N} \models \phi$. Thus, Gödel's sentence is a true Π^0_1 sentence in the form $\forall x \phi(x)$ such that $T \nvdash \forall x \phi(x)$ but '$T \vdash \phi(\bar{n})$' holds for any $n \in \omega$.

Secondly, the properties of **Truth** are essentially different from that of **Prov**. Before Gödel's work, it was thought that **Truth** = **Prov**. Thus, Gödel's first incompleteness theorem (G1) reveals the difference between the notion of provability in **PA** and the notion of truth in the standard model of arithmetic $\mathfrak{N}$. There are some differences between **Truth** and **Prov**:

- **Prov** $\subsetneq$ **Truth**, i.e., there is a true arithmetic sentence which is unprovable in **PA**;
- Tarski proves that: **Truth** is not definable in $\mathfrak{N}$ but **Prov** is definable in $\mathfrak{N}$;
- **Truth** is not arithmetic but **Prov** is recursive enumerable.

However, both **Truth** and **Prov** are not recursive and not representable in **PA**. For more details on **Truth** and **Prov**, we refer to [97, 125].

Thirdly, the differences between **Truth** and **Prov** can also be expressed in terms of *arithmetical interpretations*, defined as follows.

DEFINITION 3.2 (Arithmetical interpretations). A mapping from the set of all modal propositional variables to the set of $L(\mathbf{PA})$-sentences is called an *arithmetical interpretation* .

Every arithmetical interpretation f is uniquely extended to the mapping f^* from the set of all modal formulas to the set of $L(T)$-sentences so that f^* satisfies the following conditions:

- $f^*(p) = f(p)$ for each propositional variable p;
- f^* commutes with every propositional connective;
- $f^*(\Box A)$ is $\mathbf{Prov}_T(\ulcorner f^*(A) \urcorner)$ for every modal formula A.

In the following, we equate arithmetical interpretations f with their unique extensions f^* defined on the set of all modal formulas. In this way, Solovay's Arithmetical Completeness Theorems for **GL** and **GLS** characterize the difference between **Prov** and **Truth** via provability logic.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

THEOREM 3.3 (Solovay [122]).

Arithmetical Completeness Theorem for GL *Let T be a Σ^0_1-sound r.e. extension of* $\mathbf{Q}$. *For any modal formula ϕ in $L(\mathbf{GL})$, $\mathbf{GL} \vdash \phi$ if and only if $T \vdash f(\phi)$ for every arithmetic interpretation f.*
Arithmetical Completeness Theorem for GLS *For any modal formula ϕ, $\mathbf{GLS} \vdash \phi$ if and only if $\mathfrak{N} \models f(\phi)$ for every arithmetic interpretation f.*

Finally, one can study the notion of 'proof predicate' as given by $\mathbf{Proof}_T(x,y)$ in an abstract setting, namely as follows. Recall that T is a recursively axiomatizable consistent extension of $\mathbf{Q}$. We introduce general notions of proof predicate and provability predicate which generalize the proof predicate $\mathbf{Proof}_T(x,y)$ and the provability predicate $\mathbf{Prov}_T(x)$ defined above in Gödel's proof of G1.

DEFINITION 3.4 (Proof predicate). We say a formula $\mathbf{Prf}_T(x,y)$ is a *proof predicate* of T if it satisfies the following conditions:[9]

- $\mathbf{Prf}_T(x,y)$ is $\Delta^0_1(\mathbf{PA})$;[10]
- $\mathbf{PA} \vdash \forall x(\mathbf{Prov}_T(x) \leftrightarrow \exists y \mathbf{Prf}_T(x,y))$;
- for any $n \in \omega$ and formula ϕ, $\mathbb{N} \models \mathbf{Proof}_T(\ulcorner\phi\urcorner, \overline{n}) \leftrightarrow \mathbf{Prf}_T(\ulcorner\phi\urcorner, \overline{n})$;
- $\mathbf{PA} \vdash \forall x \forall x' \forall y(\mathbf{Prf}_T(x,y) \wedge \mathbf{Prf}_T(x',y) \rightarrow x = x')$.

DEFINITION 3.5 (Provability and consistency). We define the provability predicate $\mathbf{Pr}_T(x)$ from a proof predicate $\mathbf{Prf}_T(x,y)$ by $\exists y\, \mathbf{Prf}_T(x,y)$, and the consistency statement $\mathbf{Con}(T)$ from a provability predicate $\mathbf{Pr}_T(x)$ by $\neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$.

The items **D1**– **D3** below are called the *Hilbert–Bernays–Löb derivability conditions*. Note that **D1** holds for any provability predicate $\mathbf{Pr}_T(x)$.

DEFINITION 3.6 (Standard proof predicate). We say that provability predicate $\mathbf{Pr}_T(x)$ is *standard* if it satisfies **D2** and **D3** as follows.

D1: If $T \vdash \phi$, then $T \vdash \mathbf{Pr}_T(\ulcorner\phi\urcorner)$;
D2: If $T \vdash \mathbf{Pr}_T(\ulcorner\phi \rightarrow \varphi\urcorner) \rightarrow (\mathbf{Pr}_T(\ulcorner\phi\urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner\varphi\urcorner))$;
D3: $T \vdash \mathbf{Pr}_T(\ulcorner\phi\urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner\mathbf{Pr}_T(\ulcorner\phi\urcorner)\urcorner)$.

We say that $\mathbf{Prf}_T(x,y)$ is a *standard proof predicate* if the induced provability predicate from it is standard.

The previous definition leads to another blanket caveat:
Unless stated otherwise, we always assume that $\mathbf{Pr}_T(x)$ is a standard provability predicate, and $\mathbf{Con}(T)$ is the canonical consistency statement defined as $\neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$ via the standard provability predicate $\mathbf{Pr}_T(x)$.

[9] We can say that each proof predicate represents the relation "y is the code of a proof in T of a formula with Gödel number x".

[10] We say a formula ϕ is $\Delta^0_1(\mathbf{PA})$ if there exists a Σ^0_1 formula α such that $\mathbf{PA} \vdash \phi \leftrightarrow \alpha$, and there exists a Π^0_1 formula β such that $\mathbf{PA} \vdash \phi \leftrightarrow \beta$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

3.2.4. Properties of G2. In this section, we discuss some (sometimes subtle) comments on G2.

First of all, we examine a somewhat delicate mistake in the argument which claims that, by an easy application of the compactness theorem, we can show that for any recursive axiomatization of a consistent theory T, T cannot prove its own consistency. Visser presents this argument in [136] as an interesting dialogue between Alcibiades and Socrates:

> Suppose a consistent theory T can prove its own consistency under some axiomatization. By compactness theorem, there must be a finitely axiomatized sub-theory S of T such that S already proves the consistency of T. Since S proves the consistency of T, it must also prove the consistency of S. So, we have a finitely axiomatized theory which proves its own consistency. But G2 applies to the finite axiomatization and we have a contradiction. It follows that T cannot prove its own consistency.

The mistake in this argument is: from the fact that S can prove the consistency of T we cannot infer that S can prove the consistency of S. Some may argue that since S is a sub-theory of T and S can prove the consistency of T, then of course S can prove the consistency of S.

However, as Visser correctly points out in [136], we should carefully distinguish three perspectives of the theory T: our external perspective, the internal perspective of S, and the internal perspective of T. From each perspective, the consistency of the whole theory implies the consistency of its sub-theory. From T's perspective, S is a sub-theory of T. But from S's perspective, S may not be a sub-theory of T. From the fact that T knows that S is a sub-theory of T, we cannot infer that S also knows that S is a sub-theory of T since S is a finite sub-theory of T and may not know any information that T knows, leading to the following (dramatic) conclusion:

> *the sub-theory relation between theories is not absolute.*

Similarly, the notion of consistency is not absolute. For example, let $S = \mathbf{PA} + \neg\mathbf{Con}(\mathbf{PA})$. From G2, S is consistent from the external perspective. But since $S \vdash \neg\mathbf{Con}(S)$, the theory S is not consistent from the internal perspective of S. Note that $\mathbf{PA} \vdash \mathbf{Pr}_{\mathbf{PA}}(\mathbf{0} \neq \mathbf{0}) \to \mathbf{Pr}_{\mathbf{PA}}(\mathbf{Pr}_{\mathbf{PA}}(\mathbf{0} \neq \mathbf{0}) \to \mathbf{0} \neq \mathbf{0})$. Thus, a theory may be consistent from the external perspective but inconsistent from the internal perspective.

From Gödel's proof of G2, we cannot infer that if T is a consistent r.e. extension of $\mathbf{Q}$, then $\mathbf{Con}(T)$ is independent of T. The key point is: it is not enough to show that $T \nvdash \neg\mathbf{Con}(T)$ only assuming that T is consistent. However, we can show that $\mathbf{Con}(T)$ is independent of T assuming that T is 1-consistent.[11] In fact, the formalized version of "if T is consistent, then $\mathbf{Con}(T)$ is independent of T" is not provable in T.[12]

[11] It is an easy fact that if T is 1-consistent and S is not a theorem of T, then $\mathbf{Pr}_T(\ulcorner S \urcorner)$ is not a theorem of T.

[12] See [14, Theorem 4, p. 97] for a modal proof in **GL** of this fact using the Arithmetic Completeness Theorem for **GL**.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

DEFINITION 3.7 (Reflexivity).

- A first-order theory T containing **PA** is said to be *reflexive* if $T \vdash \mathbf{Con}(S)$ for each finite sub-theory S of T where $\mathbf{Con}(S)$ is similarly defined as $\mathbf{Con}(\mathbf{PA})$.
- We say the theory T is *essentially reflexive* if any consistent extension of T in $L(T)$ is reflexive.
- Let $\mathbf{Con}(T) \restriction x$ denote the finite consistency statement "there are no proofs of contradiction in T with $\leq x$ symbols".

Mostowski proves that **PA** is essentially reflexive (see [97, Theorem 2.6.12]). In fact one can show that for every $n \in \mathbb{N}$, $I\Sigma^0_{n+1} \vdash \mathbf{Con}(I\Sigma^0_n)$.[13] For a large class of natural theories U, Pudlák [109] shows that the lengths of the shortest proofs of $\mathbf{Con}(U) \restriction n$ for $n \in \omega$ in the theory U itself are bounded by a polynomial in n. Pudlák conjectures [109] that U does not have polynomial proofs of the finite consistency statements $\mathbf{Con}(U + \mathbf{Con}(U)) \restriction n$ for $n \in \omega$.

Finally, a big open question about G2 is: can we find a genuinely self-reference free proof of G2? As far as we know, at present there is no convincing essentially self-reference-free proofs of either G2 or of Tarski's Theorem of the Undefinability of Truth. In [135], Visser gives a self-reference-free proof of G2 from Tarski's Theorem of the Undefinability of Truth, which is a step in a program to find self-reference-free proofs of both G2 and Tarski's Theorem (see [135]). Visser's argument in [135] is model-theoretic and the main tool is the Interpretation Existence Lemma.[14] Visser's proof in [135] is not constructive. An interesting question is then whether Visser's argument can be made constructive.

3.3. *Proofs of* G1 *and* G2 *from mathematical logic.* In this section, we discuss various different proofs of G1 and G2. We mention Jech's [63] short proof of G2 for **ZF**: if **ZF** is consistent, then it is unprovable in **ZF** that there exists a model of **ZF**. Jech's proof uses the Completeness Theorem, and also yields G2 for **PA** (see [63]). Other (lengthier) proofs are discussed in Sections 3.3.1–3.3.5.

3.3.1. Rosser's proof. Rosser [111] proves a "stronger" version of G1, called Rosser's first incompleteness theorem, which only assumes the consistency of T: if T is a consistent r.e. extension of **Q**, then T is incomplete. Gödel's proof of G1 assumes that T is ω-consistent. Note that ω-consistency implies consistency. But the converse does not hold and the notion of ω-consistency is stronger than consistency since we can find examples of theories that are consistent but not ω-consistent.[15] Rosser's proof is constructive and algorithmically constructs the Rosser sentence that is independent of T. Gödel's proof of G1 uses a standard provability

[13] For a proof of this result, we refer to Hájek and Pudlák [53].

[14] We refer to [134] for more details about the Interpretation Existence Lemma.

[15] For example, assuming **PA** is consistent, then $\mathbf{PA} + \neg\mathbf{Con}(\mathbf{PA})$ is consistent, but not ω-consistent.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

predicate but Rosser's proof of G1 uses a Rosser provability predicate which is a kind of *non-standard* provability predicate, giving rise to the following.

DEFINITION 3.8. Let T be a recursively axiomatizable consistent extension of $\mathbf{Q}$, and $\mathbf{Prf}_T(x,y)$ be any proof predicate of T. Define the Rosser provability predicate $\mathbf{Pr}_T^R(x)$ to be the formula $\exists y(\mathbf{Prf}_T(x,y) \wedge \forall z \leq y \neg \mathbf{Prf}_T(\dot{\neg} x, z))$ where $\dot{\neg}$ is a function symbol expressing a primitive recursive function calculating the code of $\neg\phi$ from the code of ϕ. The fixed point of the predicate $\neg\mathbf{Pr}_T^R(x)$ is called the Rosser sentence of $\mathbf{Pr}_T^R(x)$, i.e., a sentence θ satisfying $\mathbf{PA} \vdash \theta \leftrightarrow \neg\mathbf{Pr}_T^R(\ulcorner\theta\urcorner)$.

In general, one can show that each Rosser sentence based on any Rosser provability predicate of T is independent of T. In particular, this independence does not rely on the choice of the proof predicate.

3.3.2. Recursion-theoretic proofs. Gödel's first incompleteness theorem (G1) is well-known in the context of recursion theory. Recall that $W_e = \{n \in \omega : \phi_e(n)\downarrow\}$. Let $\langle W_e : e \in \omega\rangle$ be the list of recursive enumerable subsets of $\mathbb{N}$. The following is an example of an 'effective' version of G1 (see [30]):

> there exists a recursive function f such that for any $e \in \omega$, if $W_e \subseteq$ **Truth**, then $f(e)$ is defined and $f(e) \in$ **Truth** $\setminus W_e$.

Similarly, Avigad [4] proves G1 and G2 in terms of the undecidability of the halting problem (see Theorems 3.1 and 3.2 in [4]). Another related result due to Kleene is as follows.

THEOREM 3.9 (Kleene's theorem [114, Theorem 2.2]). *For any consistent r.e. theory T that contains $\mathbf{Q}$, there exists some $t \in \omega$ such that $\varphi_t(t)\uparrow$ holds but $T \nvdash$ "$\varphi_t(t)\uparrow$".*

Kleene's proof of his theorem uses recursion theory, and is not constructive. Salehi and Seraji [114] show that there is a constructive proof of Kleene's theorem, but this constructive proof does not have the Rosser property. Salehi and Seraji [114] comment that there could be a 'Rosserian' version of this constructive proof of Kleene's theorem.

3.3.3. Proofs based on Arithmetic Completeness. Hilbert and Bernays [60] present the *Arithmetic Completeness Theorem* expressing that any recursively axiomatizable consistent theory has an arithmetically definable model. Later, Kreisel [80] and Wang [138] adapt the Arithmetic Completeness Theorem and use paradoxes to obtain undecidability results.

Now, the Arithmetic Completeness Theorem is an important tool in model-theoretic proofs of the incompleteness theorems. For more details, we refer to [67, 79, 91]. Walter Dean [28] gives a detailed discussion on how the Arithmetized Completeness Theorem provides a tool for obtaining formal incompleteness results from some certain paradoxes.

THEOREM 3.10 (Arithmetic Completeness [68, Theorem 3.1]). *Let T be a recursively axiomatized consistent extension of $\mathbf{Q}$. There exists a formula $\mathbf{Tr}_T(x)$ in $L(\mathbf{PA})$ that defines a model of T in $\mathbf{PA} + \mathbf{Con}(T)$.*

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Lemma 3.11 is a corollary of the Arithmetized Completeness Theorem, and is essential for model-theoretic proofs of the incompleteness theorems.

LEMMA 3.11 [68, 74]. *Let T be a recursively axiomatized consistent extension of* $\mathbf{Q}$, *and* $\mathbf{Tr}_T(x)$ *is the formula as asserted in Theorem 3.10. For any model M_0 of* $\mathbf{PA}+\mathbf{Con}(T)$, *there exists a model M_1 of T such that for any sentence ϕ, $M_0 \models \mathbf{Tr}_T(\ulcorner\phi\urcorner)$ if and only if $M_1 \models \phi$.*

Kreisel first applies the Arithmetized Completeness Theorem to establish model-theoretic proofs of G2 (cf. Kreisel [82], Smoryński [119], and Kikuchi [68]). Kikuchi–Tanaka [74], Kikuchi [68, 69], and Kotlarski [77] use the Arithmetized Completeness Theorem to give model-theoretic proofs of G2. For example, Kikuchi [68] proves G2 model theoretically via the Arithmetized Completeness Theorem (Lemma 3.11): if **PA** is consistent, then $\mathbf{Con}(\mathbf{PA})$ is not provable in **PA** (see [68, Theorem 3.4]).[16]

Proofs of G2 by Kreisel [82] and Kikuchi [68] do not directly yield the formalized version of G2. Kikuchi's proof of G2 in [69] is not formalizable in **PRA**. Kikuchi and Tanaka [74] prove in $\mathbf{WKL_0}$ that $\mathbf{Con}(\mathbf{PA})$ implies $\neg\mathbf{Pr}_{\mathbf{PA}}(\ulcorner\mathbf{Con}(\mathbf{PA})\urcorner)$, since the Completeness Theorem is provable in $\mathbf{WKL_0}$, and the key Lemma 3.11 used in Kikuchi's proof [68] is provable in $\mathbf{RCA_0}$.[17] Using Theorem 2.22, Tanaka [74] proves the formalized version of G2: $\mathbf{PRA} \vdash \mathbf{Con}(\mathbf{PA}) \rightarrow \mathbf{Con}(\mathbf{PA}+\neg\mathbf{Con}(\mathbf{PA}))$.

One can give a simple proof of G1 via the Diagnolisation Lemma (see [97]). Kotlarski [77] proves the formalized version of G1 and G2 via model-theoretic arguments (e.g., using the Arithmetized Completeness Theorem and some quickly growing functions). Kotlarski [77] proves the following version of G1 assuming that **PA** is ω-consistent, and shows that the following sentence is provable in **PA**:

$$\begin{aligned}&\text{if } \forall\varphi,x\{[\varphi \in \Delta_0 \wedge \forall y\mathbf{Pr}_{\mathbf{PA}}(\neg\varphi(S^x0,S^y0))] \rightarrow \neg\mathbf{Pr}_{\mathbf{PA}}(\exists y\varphi(S^x0,y))\},\\ &\text{then } \exists\varphi \in \Delta_0\exists x[\neg\mathbf{Pr}_{\mathbf{PA}}(\exists y\varphi(S^x0,y)) \wedge \neg\mathbf{Pr}_{\mathbf{PA}}(\neg\exists y\varphi(S^x0,y))].\end{aligned}$$

However, it is unknown whether the method in [77] can also give a new proof of Rosser's first incompleteness theorem. Kotlarski [77] proves the following formalized version of G2: $\mathbf{PA} \vdash \mathbf{Con}(\mathbf{PA}) \rightarrow \mathbf{Con}(\mathbf{PA}+\neg\mathbf{Con}(\mathbf{PA}))$. Later, Kotlarski [78] transforms the proof of the formalized version of G2 in [77] to a proof-theoretic version without the use of the Arithmetized Completeness Theorem.

3.3.4. Proofs based on Kolmogorov complexity. Intuitively, *Kolmogorov complexity* is a measure of the quantity of information in finite objects. Roughly speaking, the Kolmogorov complexity of a number n, denoted by $K(n)$, is the size of a program which generates n.

[16]The idea of the proof is: assuming that **PA** is consistent and $\mathbf{PA} \vdash \mathbf{Con}(\mathbf{PA})$, then we get a contradiction from the fact that there is a model M of **PA** such that $M \models \mathbf{Con}(\mathbf{PA})$.

[17]The theory $\mathbf{RCA_0}$ (Recursive Comprehension) is a subsystem of Second-Order Arithmetic. For the definition of $\mathbf{RCA_0}$, we refer to [117].

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

DEFINITION 3.12 (Kolmogorov–Chaitin Complexity [114]). For any natural number $n \in \omega$, the *Kolmogorov complexity* for n, denoted by $K(n)$, is defined as $\min\{i \in \omega \mid \varphi_i(0)\downarrow = n\}$.

If $n \leq K(n)$, then n is called random. Kolmogorov shows in 1960s that the set of non-random numbers is recursively enumerable but not recursive (cf. Odifreddi [100]). Relations between G1 and Kolmogorov complexity have been intensively discussed in the literature (cf. Li and Vitányi [90]). Chaitin [19] gives an information-theoretic formulation of G1, and proves the following weaker version of G1 in terms of Kolmogorov complexity.

THEOREM 3.13 (Chaitin [19, 114]). *For any consistent r.e. extension T of* **Q**, *there exists a constant $c_T \in \mathbb{N}$ such that for any $e \geq c_T$ and any $w \in \mathbb{N}$ we have $T \nvdash$ "$K(w) > e$".*

Salehi and Seraji [114] show that we can algorithmically construct the Chaitin constant c_T in Theorem 3.13. I.e., for a given consistent r.e. extension T of **Q**, one can algorithmically construct a constant $c_T \in \mathbb{N}$ such that for all $e \geq c_T$ and all $w \in \mathbb{N}$, we have $T \nvdash$ "$K(w) > e$" (see Theorem 3.4 in [114]). From Theorem 3.13, it is not clear whether "$K(w) > e$" holds (or whether "$K(w) > e$" is independent of T). Salehi and Seraji [114] show that Chaitin's proof of G1 is non-constructive: there is no algorithm such that given any consistent r.e. extension T of **Q** we can compute some w_T such that $K(w_T) > c_T$ holds where c_T is the Chaitin constant we can compute as in Theorem 3.13 (see [114, Theorem 3.5]). If such an algorithm exists, then for any consistent r.e. extension T of **Q**, we can compute some c_T and w_T such that $K(w_T) > c_T$ is true but unprovable in T.

Salehi and Seraji [114] also strengthen Chaitin's Theorem 3.13 assuming T is Σ^0_1-sound: if T is a Σ^0_1-sound r.e. theory extending **Q**, then there exists some c_T (which is computable from T) such that for any $e \geq c_T$ there are cofinitely many w's such that "$K(w) > e$" is independent of T (see [114, Corollary 3.7]). Using a version of the Pigeonhole Principle in **Q**, Salehi and Seraji [114] also prove the Rosserian form of Chaitin's Theorem: for any consistent r.e. extension T of **Q**, there is a constant c_T (which is computable from T) such that for any $e \geq c_T$ there are cofinitely many w's such that "$K(w) > e$" is independent of T (see [114, Theorem 3.9]).

Kikuchi [69] proves the following formalized version of G1 via Kolmogorov complexity for any consistent r.e. extension T of **PA**: there exists $e \in \omega$ with:

- $T \vdash \mathbf{Con}(T) \to \forall x(\neg \mathbf{Pr}_T(\ulcorner K(x) > e \urcorner))$;
- $T \vdash \omega\text{-}\mathbf{Con}(T) \to \forall x(e < K(x) \to \neg \mathbf{Pr}_T(\ulcorner K(x) \leq e \urcorner))$.

However, this proof is not constructive. Moreover, Kikuchi [69] proves G2 via Kolmogorov complexity and the Arithmetic Completeness Theorem: if T is a consistent r.e. extension of **PA**, then $T \nvdash \mathbf{Con}(T)$. Kikuchi's proof of G2 in [69] cannot be formalized in **PRA** but can be carried out within $\mathbf{WKL_0}$. Thus we can also obtain a formalized version of G2 in $\mathbf{WKL_0}$ by Theorem 2.22.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

3.3.5. Model-theoretic proofs. Adamowicz and Bigorajska [1] prove G2 via model-theoretic method using the notion of 1-closed models and existentially closed models.

DEFINITION 3.14.

- A model M of a theory T is called 1-closed (w.r.t. T) if for any $a_1, \ldots, a_n$ in M, any Σ_1 formula ϕ, and any M' such that $M \prec_0 M'$ and $M' \models T$, we have: if $M' \models \phi(a_1, \cdots, a_n)$, then $M \models \phi(a_1, \ldots, a_n)$. In other words, we can say that M is 1-closed if for any M' such that $M \prec_0 M'$, we have $M \prec_1 M'$.
- Let $\mathcal{K}$ be a class of structures in the same language. A model $M \in \mathcal{K}$ is *existentially closed* in $\mathcal{K}$ if for every model $N \supseteq M$ such that $N \in \mathcal{K}$, we have $M \preceq_1 N$: every existential formula with parameters from M which is satisfied in N is already satisfied in M.

Adamowicz and Bigorajska [1] first prove G2 without the use of the Arithmetized Completeness Theorem: every 1-closed model of any subtheory T of **PA** extending $I\Delta_0 + \mathbf{exp}$ satisfies $\neg\mathbf{Con}(\mathbf{PA})$. Then Adamowicz and Bigorajska [1] prove the formalized version of G2 via the idea of existentially closed models and the Arithmetized Completeness Theorem: $\mathbf{PA} \vdash \mathbf{Con}(\mathbf{PA}) \rightarrow \mathbf{Con}(\mathbf{PA} + \neg\mathbf{Con}(\mathbf{PA}))$ (see [1, Theorem 2.1]). This is proved by showing that an arbitrary model of $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$ satisfies $\mathbf{Con}(\mathbf{PA} + \neg\mathbf{Con}(\mathbf{PA}))$.

3.4. *Proofs of* G1 *and* G2 *based on logical paradox.* We provide a survey of proofs of incompleteness theorems based on 'logical paradox'.

3.4.1. Introduction. As noted in Section 3.2.1, Gödel's incompleteness theorems are closely related to paradox and self-reference. In fact, Gödel comments in his famous paper that "any epistemological antinomy could be used for a similar proof of the existence of undecidable propositions".

Now, the *Liar Paradox* is an old and most famous paradox in modern science. In Gödel's proof of G1, we can view Gödel's sentence as the formalization of the Liar Paradox. Gödel's sentence concerns the notion of provability but the liar sentence in the Liar Paradox concerns the notion of truth in the standard model of arithmetic. There is a big difference between the notion of provability and truth. Gödel's sentence does not lead to a contradiction as the Liar sentence does.

Besides the Liar Paradox, many other paradoxes have been used to give new proofs of incompleteness theorems: for example, Berry's Paradox in [13, 19, 68, 73, 74, 137], Grelling–Nelson's Paradox in [25], the Unexpected Examination Paradox in [36, 83], and Yablo's Paradox in [26, 70, 85, 106]. We now discuss some of these paradoxes in detail.

3.4.2. Berry's Paradox. Berry's Paradox introduced by Russell [112] is the paradox that "the least integer not nameable in fewer than 19 syllables" is itself a name consisting of 18 syllables. Informally, we say that an expression names a natural number n if n is the unique natural number satisfying the expression. Berry's Paradox can be formalized in formal systems by

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

interpreting the concept of "name" suitably. The following is Boolos's formulation of the concept of "name" in [13].

DEFINITION 3.15 (Boolos [13]). Let $n \in \omega$ and $\varphi(x)$ be a formula with only one free variable x. We say that $\varphi(x)$ names n if $\mathfrak{N} \models \varphi(\overline{n}) \wedge \forall v_0 \forall v_1 (\varphi(v_0) \wedge \varphi(v_1) \rightarrow v_0 = v_1)$.

Proofs of the incompleteness theorems based on Berry's Paradox have been given by Vopěnka [137], Chaitin [19], Boolos [13], Kikuchi–Kurahashi–Sakai [73], Kikuchi [68], and Kikuchi–Tanaka [74]. In fact, Robinson first uses Berry's Paradox in [110] to prove Tarski's theorem on the undefinability of truth, which anticipates the later use of Berry's Paradox to obtain incompleteness results by Vopěnka [137], Boolos [13], and Kikuchi [69].

Boolos [13] proves a weak form of G1 in the 1980s by formalizing Berry's Paradox in arithmetic via considering the length of formulas that name natural numbers in the standard model of arithmetic. Using this formulation of the concept of "name", Boolos [13] first shows that Berry's Paradox leads to a proof of G1 in the following form: there is no algorithm whose output contains all true statements of arithmetic and no false ones (i.e., the theory of true arithmetic is not recursively axiomatizable). Barwise praises Boolos's proof as "very lovely and the most straightforward proof of Gödel's incompleteness theorem that I have ever seen". The optimal sufficient and necessary condition for the independence of a Boolos sentence from **PA** is that $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$ is consistent (see [114]).

Boolos's theorem is different from Gödel's theorem in the following way:

- Boolos's theorem refers to the concept of truth but Gödel's theorem does not;
- Boolos's proof is not constructive, and we can prove that there is no algorithm for computing the true but unprovable sentence;
- Boolos's theorem is weaker than Gödel's first incompleteness theorem, and hence we cannot obtain the second incompleteness theorem from Boolos's theorem in the standard way (see [73]).

Boolos's proof is modified by Kikuchi and Tanaka in [68, 74]. The difference between Kikuchi's proof and Boolos's proof lies in the interpretation of the word "name". Kikuchi [68] modifies Boolos's formulation of the concept of "name" by replacing "truth" with "provability" in the definition.

DEFINITION 3.16 (Kikuchi [68, Definition 3.1]). Let $n \in \omega$ and $\varphi(x)$ be a formula with only one free variable x. We say that $\varphi(x)$ names n if $\mathbf{PA} \vdash \varphi(\overline{n}) \wedge \forall v_0 \forall v_1 (\varphi(v_0) \wedge \varphi(v_1) \rightarrow v_0 = v_1)$.

Using this formulation of the concept of "name", Kikuchi [68] gives a proof-theoretic proof of G1 by formalizing Berry's Paradox without the use of the Diagnolisation Lemma. Kikuchi [68] constructs a sentence θ and shows that if **PA** is consistent, then $\neg\theta$ is not provable in **PA**; if **PA** is

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

ω-consistent, then θ is not provable in **PA** (see [68, Theorem 2.2]). Note that Kikuchi's proof of G1 in [68] is constructive. Kikuchi and Tanaka [74] reformulate Kikuchi's proof of G1 in [68], and show in $\mathbf{WKL_0}$ that if $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$ is consistent, then θ is independent of **PA**. By Theorem 2.22, Kikuchi and Tanaka [74] prove the formalized version of G1: $\mathbf{PRA} \vdash \mathbf{Con}(\mathbf{PA} + \mathbf{Con}(\mathbf{PA})) \rightarrow \neg\mathbf{Pr}_{\mathbf{PA}}(\ulcorner\theta\urcorner) \wedge \neg\mathbf{Pr}_{\mathbf{PA}}(\ulcorner\neg\theta\urcorner)$. An interesting question not covered in [68, 74] is whether we can improve Kikuchi's proof of G1 by only assuming that **PA** is consistent.

Vopěnka [137] proves G2 for **ZF** by formalizing Berry's Paradox, via adopting Kikuchi's definition of the concept of "name" in [68] over models of **ZF**[18]: $\mathbf{Con}(\mathbf{ZF})$ is not provable in **ZF**. Vopěnka's proof uses the Completeness Theorem but does not use the Arithmetic Completeness Theorem. Kikuchi, Kurahashi, and Sakai [73] show that Vopěnka's method can be adapted to prove G2 for **PA** based on Kikuchi's formalization of Berry's Paradox in [68] with an application of the Arithmetic Completeness Theorem.

Proofs of G1 and G2 based on Berry's Paradox by Vopěnka [137], Chaitin [19], Boolos [13], and Kikuchi [68] do not use the Diagnolisation Lemma. We can also prove G1 based on Berry's Paradox using the Diagnolisation Lemma. For example, Kikuchi, Kurahashi, and Sakai [73] adopt Kikuchi's definition of the concept of "name" in [68], and show that the independent statement in Kikuchi's proof in [68] can be obtained by using the Diagnolisation Lemma.

In summary, the distinctions between using and not using the Diagnolisation Lemma, and between using and not using the Arithmetic Completeness Theorem are not essential for proofs of G1 and G2 based on Berry's Paradox. From the above discussions, we can characterize different proofs of G1 and G2 based on Berry's Paradox by the method of interpreting the word "name": Boolos [13] uses the standard model of arithmetic; Kikuchi [68] uses provability in arithmetic; Chaitin [19] and Kikuchi [69] use Kolmogorov complexity; Kikuchi and Tanaka [74] use nonstandard models of arithmetic; and Vopěnka [137] uses models of **ZF** (see [73]).

3.4.3. Unexpected Examination and Grelling–Nelson's Paradox. First of all, Kritchman and Raz [83] give a new proof of G2 based on Chaitin's incompleteness theorem and an argument that resembles the Unexpected Examination Paradox[19] (for more details, we refer to [83]): for any consistent r.e. extension T of **PA**, if T is consistent, then $T \nvdash \mathbf{Con}(T)$.

Secondly, we say a one-place predicate is "heterological" if it does not apply to itself (e.g., "long" is heterological, since it's not a long expression).

[18]I.e., we say that $\varphi(x)$ names n in **ZF** if $\mathbf{ZF} \vdash \varphi(\overline{n}) \wedge \forall v_0 \forall v_1 (\varphi(v_0) \wedge \varphi(v_1) \rightarrow v_0 = v_1)$ where $\varphi(x)$ is a formula with only one free variable x (see [137]).

[19]The Unexpected Examination Paradox is formulated as follows in [83]. The teacher announces in class: "next week you are going to have an exam, but you will not be able to know on which day of the week the exam is held until that day". The exam cannot be held on Friday, because otherwise, the night before the students will know that the exam is going to be held the next day. Hence, in the same way, the exam cannot be held on Thursday. In the same way, the exam cannot be held on any of the days of the week.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Consider the question: is the predicate "heterological" we have just defined heterological? If "heterological" is heterological, then it isn't heterological; and if "heterological" isn't heterological, then it is heterological. This contradiction is called Grelling–Nelson's Paradox.

Cieśliński [25] presents semantic proofs of G2 for **ZF** and **PA** based on Grelling–Nelson's Paradox. For a theory T containing **ZF**, Cieśliński defines the sentence $\mathbf{HET_T}$ which says intuitively that the predicate "heterological" is itself heterological, and then shows that $T \nvdash \mathbf{HET_T}$ and $T \vdash \mathbf{HET_T} \leftrightarrow \mathbf{Con}(T)$. Finally, Cieśliński shows how to adapt the proof of G2 for **ZF** to a proof of G2 for **PA**. In fact, Cieśliński [25] proves the semantic version of G2: if T has a model, then $T + \neg\mathbf{Con}(T)$ has a model (i.e., $T \nvdash \mathbf{Con}(T)$).

3.4.4. Yablo's Paradox. We discuss proofs of G1 and G2 based on Yablo's Paradox in the literature. Yablo's Paradox is an infinite version of the Liar Paradox proposed in [146]: consider an infinite sequence $Y_1, Y_2, \ldots$ of propositions such that each Y_i asserts that Y_j are false for all $j > i$. Different proofs of G1 and G2 based on Yablo's Paradox have been given by some authors (e.g., Priest [106], Cieśliński–Urbaniak [26], and Kikuchi–Kurahashi [70]).

Recall that we assume by default that T is a consistent r.e. extension of **Q**. Priest [106] first points out that G1 can be proved by formalizing Yablo's Paradox. Priest defines a formula $Y(x)$ as follows which says that for any $y > x$, $Y(y)$ is not provable in T.

Definition 3.17 ([26, 85]). A formula $Y(x)$ is called a Yablo formula of T if $T \vdash \forall x(Y(x) \leftrightarrow \forall y > x \neg\mathbf{Pr}_T(\ulcorner Y(\dot{y})\urcorner))$.

Cieśliński and Urbaniak originally prove the following version of G1, and show that each instance $Y(\overline{n})$ of the Yablo formula is independent of T if T is Σ^0_1-sound (or 1-consistent).

Theorem 3.18 ([26, Theorem 19]; see also [85, Theorem 4]). *Let $Y(x)$ be a Yablo formula.*

- *If T is consistent, then $T \nvdash Y(\overline{n})$.*
- *If T is Σ^0_1-sound, then $T \nvdash \neg Y(\overline{n})$.*

Cieśliński and Urbaniak originally prove that $T \vdash \forall x(Y(x) \leftrightarrow \mathbf{Con}(T))$ (see [26, Theorems 21 and 22]). As a corollary, we have $T \vdash \forall x \forall y(Y(x) \leftrightarrow Y(y))$, and G2 holds: if T is consistent, then $T \nvdash \mathbf{Con}(T)$.

Definition 3.19 ([26, 85]). A formula $Y^R(x)$ is called a Rosser-type Yablo formula of $\mathbf{Prf}_T(x,y)$ if $\mathbf{PA} \vdash \forall x(Y^R(x) \leftrightarrow \forall y > x \neg\mathbf{Pr}^R_T(x)(\ulcorner Y^R(\dot{y})\urcorner))$.

Theorem 3.20 shows that the Rosser-type Yablo formula is independent of any Σ^0_1-sound theory T.

Theorem 3.20 ([85, Theorem 10]). *Let $\mathbf{Prf}_T(x,y)$ be any standard proof predicate of T, and $Y^R(x)$ be any Rosser-type Yablo formula of $\mathbf{Prf}_T(x,y)$. Given $n \in \omega$, if T is consistent, then $T \nvdash Y^R(\overline{n})$; if T is Σ^0_1-sound, then $T \nvdash \neg Y^R(\overline{n})$.*

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

The independence of $Y^R(\overline{n})$ for T which is not Σ^0_1-sound is discussed in [85]. For a consistent but not Σ^0_1-sound theory, the situation of Rosser-type Yablo formulas is quite different from that of Rosser sentences. Kurahashi [85] shows that for any consistent but not Σ^0_1-sound theory, the independence of each instance of a Rosser-type Yablo formula depends on the choice of standard proof predicates (see Theorems 12 and 25 in [85]). Kurahashi [85] shows that for any consistent but not Σ^0_1-sound theory T, there is a standard proof predicate of T such that each instance $Y^R(\overline{n})$ of the Rosser-type Yablo formula $Y^R(x)$ based on this proof predicate is provable in T for any $n \in \omega$. Moreover, Kurahashi [85] constructs a standard proof predicate of T and a Rosser-type Yablo formula $Y^R(x)$ based on this proof predicate such that each instance of $Y^R(x)$ is independent of T. Proofs of these results use the technique of Guaspari and Solovay in [50].

Cieśliński and Urbaniak [26] conjecture that any two distinct instances $Y^R(\overline{m})$ and $Y^R(\overline{n})$ of a Rosser-type Yablo formula $Y^R(x)$ based on a standard proof predicate are not provably equivalent. Leach–Krouse [84] and Kurahashi [85] construct a standard proof predicate, and a Rosser-type Yablo formula $Y^R(x)$ based on this proof predicate such that $T \vdash \forall x \forall y (Y^R(x) \leftrightarrow Y^R(y))$ (see [84, Theorem 9] and [85, Corollary 21]).

Kurahashi [85] constructs a partial counterexample to Cieśliński and Urbaniak's conjecture: a standard proof predicate, and a Rosser-type Yablo formula $Y^R(x)$ based on this proof predicate such that $\forall x \forall y (Y^R(x) \leftrightarrow Y^R(y))$ is not provable in T ([85, Corollary 20]). Thus the provability of the sentence $\forall x \forall y (Y^R(x) \leftrightarrow Y^R(y))$ also depends on the choice of standard proof predicates (see [85, Corollaries 20 and 21]). Proofs of these results by Leach-Krouse and Kurahashi also use the technique of Guaspari and Solovay in [50]. An interesting open question is: whether there is a standard proof predicate such that $Y^R(\overline{n})$ and $Y^R(\overline{n+1})$ are not provably equivalent for some $n \in \omega$ (see [85])?

3.4.5. Beyond arithmetization. All the proofs of G1 we have discussed use arithmetization. Andrzej Grzegorczyk proposes the theory **TC** in as a possible alternative theory for studying incompleteness and undecidability, and shows that **TC** is essentially incomplete and mutually interpretable with **Q** without arithmetization.

Now, in **PA** we have numbers that can be added or multiplied; while in **TC**, one has strings (or texts) that can be concatenated. In Gödel's proof, the only use of numbers is coding of syntactical objects. The motivations for accepting strings rather than numbers as the basic notion are as follows: on metamathematical level, the notion of computability can be defined without reference to numbers; for Grzegorczyk, dealing with texts is philosophically better justified since intellectual activities like reasoning, communicating, or even computing involve working with texts not with numbers. Thus, it is natural to define notions like undecidability directly in terms of texts instead of natural numbers. Grzegorczyk only proves the incompleteness of **TC** in. Later, Grzegorczyk and Zdanowski [48] prove that **TC** is essentially incomplete.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

3.5. *Concrete incompleteness.*

3.5.1. Introduction. All proofs of Gödel's incompleteness theorems we have discussed above make use of meta-mathematical or logical methods, and the independent sentence constructed has a clear meta-mathematical or logical flavour which is devoid of real mathematical content. To be blunt, from a purely mathematical point of view, Gödel's sentence is artificial and not mathematically interesting. Gödel's sentence is constructed not by reflecting about arithmetical properties of natural numbers, but by reflecting about an axiomatic system in which those properties are formalized (see [61]). A natural question is then: can we find true sentences not provable in **PA** with real mathematical content? The research program *concrete incompleteness* is the search for natural independent sentences with real mathematical content.

This program has received a lot of attention because despite Gödel's incompleteness theorems, one can still cherish the hope that all natural and mathematically interesting sentences about natural numbers are provable or refutable in **PA**, and that elementary arithmetic is complete w.r.t. natural and mathematically interesting sentences. However, after Gödel, many natural independent sentences with real mathematical content have been found. These independent sentences have a clear mathematical flavor, and do not refer to the arithmetization of syntax and provability.

In this section, we provide an overview of the research on concrete incompleteness. The survey paper [15] provides a good overview on the state-of-the-art up to Autumn 2006. For more detailed discussions about concrete incompleteness, we refer to Cheng [21], Bovykin [15], and Friedman [40].

3.5.2. Paris–Harrington and beyond. Paris and Harrington [103] propose the first mathematically natural statement independent of **PA**: the *Paris–Harrington Principle* (PH) which generalizes the *finite Ramsey theorem*. Gödel's sentence is a pure logical construction (via the arithmetization of syntax and provability predicate) and has no relevance with classic mathematics (without any combinatorial or number-theoretic content). On the contrary, Paris–Harrington Principle is an independent arithmetic sentence from classic mathematics with combinatorial content as we will show. We refer to [61] for more discussions about the distinction between mathematical arithmetic sentences and meta-mathematical arithmetic sentences.

Definition 3.21 (Paris–Harrington Principle (PH) [103]).

- For set X and $n \in \omega$, let $[X]^n$ be the set of all n-elements subset of X. We identify n with $\{0, \ldots, n-1\}$.
- For all $m, n, c \in \omega$, there is $N \in \omega$ such that for all $f : [N]^m \to c$, we have: $(\exists H \subseteq N)(|H| \geq n \wedge H$ is homogeneous for $f \wedge |H| > \min(H))$.

Theorem 3.22 (Paris–Harrington [103]). *The principle* PH *is true but not provable in* **PA**.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Now, PH has a clear combinatorial flavor, and is of the form $\forall x \exists y \psi(x, y)$ where ψ is a Δ_0^0 formula. It can be shown that for any given natural number n, $\mathbf{PA} \vdash \exists y \psi(\overline{n}, y)$, i.e., all particular *instances* of PH are provable in **PA**.

Following PH, many other mathematically natural statements independent of **PA** with combinatorial or number-theoretic content have been formulated: the Kanamori–McAloon principle, the Kirby–Paris sentence [104], the Hercules–Hydra game [104], the Worm principle [7, 56], the flipping principle [75], the arboreal statement [93], the kiralic and regal principles [27], and the Pudlák's principle [52, 107] (see [15, p. 40]). In fact, all these principles are equivalent to PH (see [15, p.40]).

An interesting and amazing fact is: all the above mathematically natural principles are in fact provably equivalent in **PA** to a certain meta-mathematical sentence. Consider the following reflection principle for Σ_1^0 sentences: for any Σ_1^0 sentence ϕ in $L(\mathbf{PA})$, if ϕ is provable in **PA**, then ϕ is true. Using the arithmetization of syntax, one can write this principle as a sentence of $L(\mathbf{PA})$, and denote it by $\mathsf{Rfn}_{\Sigma_1^0}(\mathbf{PA})$ (see [97, p. 301]). McAloon has shown that $\mathbf{PA} \vdash \mathsf{PH} \leftrightarrow \mathsf{Rfn}_{\Sigma_1^0}(\mathbf{PA})$ (see [97, p. 301]), and similar equivalences can be established for the other independent principles mentioned above. Equivalently, all these principles are equivalent to so-called 1-consistency of **PA** (see [8, p. 36], [7, p. 3], and [97, p. 301]).

The above phenomenon indicates that the difference between mathematical and meta-mathematical statements is perhaps not as huge as we might have expected. Moreover, the above principles are provable in fragments of Second-Order Arithmetic and are more complex than Gödel's sentence: Gödel's sentence is equivalent to $\mathbf{Con}(\mathbf{PA})$ in **PA**; but all these principles are not only independent of **PA** but also independent of $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$ (see [8, p. 36] and [97, p. 301]).

3.5.3. Harvey Friedman's contributions. Incompleteness would not be complete without mentioning the work of Harvey Friedman who is a central figure in research on the foundations of mathematics after Gödel. He has made many important contributions to concrete mathematical incompleteness. The following quote is telltale:

> the long range impact and significance of ongoing investigations in the foundations of mathematics is going to depend greatly on the extent to which the Incompleteness Phenomena touches normal concrete mathematics (see [40, p. 7]).

In the following, we give a brief introduction to H. Friedman's work on concrete mathematical incompleteness. In his early work, H. Friedman examines how one uses large cardinals in an essential and natural way in number theory, as follows.

> the quest for a simple meaningful finite mathematical theorem that can only be proved by going beyond the usual axioms for mathematics has been a goal in the foundations of mathematics since Gödel's incompleteness theorems (see [39, p. 805]).

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

H. Friedman shows in [38, 39] that there are many mathematically natural combinatorial statements in $L(\mathbf{PA})$ that are neither provable nor refutable in **ZFC** or **ZFC** + large cardinals. H. Friedman's more recent monograph [40] is a comprehensive study of concrete mathematical incompleteness. H. Friedman studies concrete mathematical incompleteness over different systems, ranging from weak subsystems of **PA** to higher-order arithmetic and **ZFC**. H. Friedman lists many concrete mathematical statements in $L(\mathbf{PA})$ that are independent of subsystems of **PA**, or stronger theories like higher-order arithmetic and set theory.

The theories $\mathbf{RCA_0}$ (Recursive Comprehension), $\mathbf{WKL_0}$ (Weak Konig's Lemma), $\mathbf{ACA_0}$ (Arithmetical Comprehension), $\mathbf{ATR_0}$ (Arithmetic Transfinite Recursion), and Π^1_1-$\mathbf{CA_0}$ (Π^1_1-Comprehension) are the most famous five subsystems of Second-Order Arithmetic (**SOA**), and are called the 'Big Five'. For the definition of **SOA** and the 'Big Five', we refer to [117].

To give the reader a better sense of H. Friedman's work, we list some sections dealing with concrete mathematical incompleteness in [40].

- Section 0.5 on Incompleteness in Exponential Function Arithmetic.
- Section 0.6 on Incompleteness in Primitive Recursive Arithmetic, Single Quantifier Arithmetic, $\mathbf{RCA_0}$, and $\mathbf{WKL_0}$.
- Section 0.7 on Incompleteness in Nested Multiply Recursive Arithmetic and Two Quantifier Arithmetic.
- Section 0.8 on Incompleteness in Peano Arithmetic and $\mathbf{ACA_0}$.
- Section 0.9 on Incompleteness in Predicative Analysis and $\mathbf{ATR_0}$.
- Section 0.10 on Incompleteness in Iterated Inductive Definitions and Π^1_1-$\mathbf{CA_0}$.
- Section 0.11 on Incompleteness in Second-Order Arithmetic and $\mathbf{ZFC}^-$.[20]
- Section 0.12 on Incompleteness in Russell Type Theory and Zermelo Set Theory.
- Section 0.13 on Incompleteness in **ZFC** using Borel Functions.
- Section 0.14 on Incompleteness in **ZFC** using Discrete Structures.

H. Friedman [40] provides us with examples of concrete mathematical theorems not provable in subsystems of Second-Order Arithmetic stronger than **PA**, and a number of concrete mathematical statements provable in Third-Order Arithmetic but not provable in Second-Order Arithmetic.

Related to Friedman's work, Cheng [21, 24] gives an example of concrete mathematical theorems based on Harrington's principle which is isolated from the proof of the Harrington's Theorem (the determinacy of Σ^1_1 games implies the existence of zero sharp), and shows that this concrete theorem saying that Harrington's principle implies the existence of zero sharp is expressible in Second-Order Arithmetic, not provable in Second-

[20] $\mathbf{ZFC}^-$ denotes **ZFC** with the Power Set Axiom deleted and Collection instead of Replacement.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Order Arithmetic or Third-Order Arithmetic, but provable in Fourth-Order Arithmetic (i.e., the minimal system in higher-order arithmetic to prove this concrete theorem is Fourth-Order Arithmetic).

Many other examples of concrete mathematical incompleteness, and the discussion of this subject in 1970s–1980s can be found in the four volumes [10, 101, 115, 116]. Weiermann's work in [139–143] provides us with more examples of naturally mathematical independent sentences. We refer to [40] for new advances in Boolean Relation Theory and for more examples of concrete mathematical incompleteness.

§4. The limit of the applicability of G1.

4.1. ***Introduction.*** In this section, we discuss the limit of the applicability of G1 based on the following two questions.

- To what extent does G1 apply to extensions of **PA**?
- To what extent does G1 apply to theories weaker than **PA** w.r.t. interpretation?

Definition 4.1 (Conservativity).

- Let Γ denote either Σ^0_n or Π^0_n for some $n \geq 1$, and Γ^d denote either Π^0_n or Σ^0_n.
- We say a sentence φ is Γ-conservative over theory T if for any Γ sentence ψ, $T \vdash \psi$ whenever $T + \varphi \vdash \psi$.

We list some generalizations of G1 needed below.

Fact 4.2 (Guaspari [49]). Let T be a consistent r.e. extension of **Q**. Then there is a Γ^d sentence ϕ such that ϕ is Γ-conservative over T and $T \nvdash \phi$.

If $T \vdash \neg\phi$, then ϕ is not Γ-conservative over T because T is consistent. Thus, we can view Fact 4.2 as an extension of Rosser's first incompleteness theorem. Solovay improves this fact and shows that there is a Γ^d sentence ϕ such that ϕ is Γ-conservative over T, $\neg\phi$ is Γ^d-conservative over T, but ϕ is independent of T.

Fact 4.3 Let $\{T_n : n \in \omega\}$ be an r.e. sequence of consistent theories extending **Q**. Then there is a Π^0_1 sentence ϕ such that for any $n \in \omega$, $T_n \nvdash \phi$, and $T_n \nvdash \neg\phi$.

4.2. ***Generalizations of*** G1 ***beyond*** **PA.** We study generalization of G1 for extensions of **PA** w.r.t. interpretation. We know that G1 applies to all consistent r.e. extensions of **PA**. A natural question is then: whether G1 can be extended to non-r.e. arithmetically definable extensions of **PA**?

Kikuchi–Kurahashi [71] and Salehi–Seraji [113] make contributions to generalize Gödel–Rosser's first incompleteness theorem to non-r.e. arithmetically definable extensions of **PA**.

Definition 4.4 ([71]). Let T be a consistent extension of **Q**.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

- T is Σ^0_n-definable if there is a Σ^0_n formula $\phi(x)$ such that n is the Gödel number of some sentence of T if and only if $\mathfrak{N} \models \phi(\overline{n})$.[21]
- T is Σ^0_n-sound if for all Σ^0_n sentences ϕ, $T \vdash \phi$ implies $\mathfrak{N} \models \phi$; T is sound if T is Σ^0_n-sound for any $n \in \omega$.
- T is Σ^0_n-consistent if for all Σ^0_n formulas ϕ with $\phi = \exists x \theta(x)$ and $\theta \in \Pi^0_{n-1}$, if $T \vdash \neg\theta(\overline{n})$ for all $n \in \omega$, then $T \nvdash \phi$.
- T is Π^0_n-decisive if for all Π^0_n sentences ϕ, either $T \vdash \phi$ or $T \vdash \neg\phi$ holds.

From G1, we have: if T is a Σ^0_1-definable and Σ^0_1-sound extension of $\mathbf{Q}$, then T is not Π^0_1-decisive. Kikuchi and Kurahashi [71] generalize G1 to arithmetically definable theories via the notion of " Σ^0_n-sound".

THEOREM 4.5 ([71, Theorem 4.8] and [113, Theorem 2.5]). *If T is a Σ^0_{n+1}-definable and Σ^0_n-sound extension of $\mathbf{Q}$, then T is not Π^0_{n+1}-decisive.*

Salehi and Seraji [113] point out that Theorem 4.5 has a constructive proof: given a Σ^0_{n+1}-definable and Σ^0_n-sound extension T of $\mathbf{Q}$, one can effectively construct a Π^0_{n+1} sentence which is independent of T. The optimality of Theorem 4.5 is shown by Salehi and Seraji in [113]: there exists a Σ^0_{n-1}-sound and Σ^0_{n+1}-definable complete extension of $\mathbf{Q}$ for any $n \geq 1$ ([113, Theorem 2.6]).

Salehi and Seraji [113] generalize G1 to arithmetically definable theories via the notion of " Σ^0_n-consistent".

THEOREM 4.6 ([71, Theorem 4.9] and [113, Theorem 4.3]). *If T is a Σ^0_{n+1}-definable and Σ^0_n-consistent extension of $\mathbf{Q}$, then T is not Π^0_{n+1}-decisive.*

Theorem 4.6 is also optimal: the complete Σ^0_{n-1}-sound and Σ^0_{n+1}-definable theory constructed in the proof of Theorem 2.6 in [113] is also Σ^0_{n-1}-consistent since if a theory is Σ^0_n-sound, then it is Σ^0_n-consistent. The proof of Theorem 4.6 cannot be constructive as the following theorem shows.

THEOREM 4.7 (Non-constructivity of Σ^0_n-consistency incompleteness [113, Theorem 4.4]). *For $n \geq 3$, there is no (partial) recursive function f (even with the oracle 0^n) such that if m codes (the Gödel code) a Σ^0_{n+1}-formula which defines an Σ^0_n-consistent extension T of $\mathbf{Q}$, then $f(m)$ halts and codes a Π^0_{n+1} sentence which is independent of T.*[22]

In summary, G1 can be generalized to the incompleteness of Σ^0_{n+1}-definable and Σ^0_n-sound extensions of $\mathbf{Q}$ constructively; and to the incompleteness of Σ^0_{n+1}-definable and Σ^0_n-consistent extensions of $\mathbf{Q}$ non-constructively (when $n > 2$).

4.3. *Generalizations of* G1 *below* PA. We study generalizations of G1 for theories weaker than **PA** w.r.t. interpretation.

[21] Recall that $\mathfrak{N}$ is the standard model of arithmetic.

[22] Salehi and Seraji [113] remark that there indeed exists some 0^{n+1}-(total) recursive function f such that if m codes a Σ^0_{n+1}-formula defining an Σ^0_n-consistent extension T of $\mathbf{Q}$, then $f(m)$ halts and codes a Π^0_{n+1} sentence independent of T.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

4.3.1. Generalizations of G1 *via interpretability.* We show that G1 can be generalized to theories weaker than **PA** via interpretability. Indeed, there exists a weak recursively axiomatizable consistent subtheory T of **PA** such that each recursively axiomatizable theory S in which T is interpretable is incomplete (see [125]). To generalize this fact further, we propose a new notion "G1 holds for T", as follows.

DEFINITION 4.8 Let T be a consistent r.e. theory. We say G1 *holds for* T if for any recursively axiomatizable consistent theory S, if T is interpretable in S, then S is incomplete.

First of all, for a consistent r.e. theory T, it is not hard to show that the followings are equivalent (see [22]):

- G1 holds for T.
- T is essentially incomplete.
- T is essentially undecidable.

It is well-known that G1 holds for many weaker theories than **PA** w.r.t. interpretation (e.g., Robinson Arithmetic **Q**).

Secondly, we mention theories weaker than **PA** w.r.t. interpretation for which G1 holds. We first review some essentially undecidable theories weaker than **PA** w.r.t. interpretation from the literature (i.e., G1 holds for these theories). For the definition of theory **Q**, $I\Sigma_n$, $B\Sigma_n$, $\mathbf{PA}^-$, $\mathbf{Q}^+$, $\mathbf{Q}^-$, $\mathbf{S}^1_2$, **AS**, **EA**, **PRA**, **R**, $\mathbf{R}_0$, $\mathbf{R}_1$, and $\mathbf{R}_2$, we refer to Section 2.

Robinson shows that any consistent r.e. theory that interprets **Q** is undecidable, and hence **Q** is essentially undecidable. The fact that **Q** is essentially undecidable is very useful and can be used to prove the essentially undecidability of other theories via Theorem 2.6. Since **Q** is finitely axiomatized, it follows that any theory that weakly interprets **Q** is also undecidable.

The Lindenbaum algebras of all r.e. theories that interpret **Q** are recursively isomorphic (see Pour-El and Kripke [105]). In fact, **Q** is minimal essentially undecidable in the sense that if deleting any axiom of **Q**, then the remaining theory is not essentially undecidable and has a complete decidable extension (see [125, Theorem 11, p. 62]).

Thirdly, Nelson [98] embarks on a program of investigating how much mathematics can be interpreted in Robinson's Arithmetic **Q**: what can be interpreted in **Q**, and what cannot be interpreted in **Q**? In fact, **Q** represents a rich degree of interpretability since a lot of stronger theories are interpretable in it as we will show in the following passages. For example, using Solovay's method of shortening cuts (see [50]), one can show that **Q** interprets fairly strong theories like $I\Delta_0+\Omega_1$ on a definable cut.

Fourth, we discuss some prominent fragments of **PA** extending **Q** from the literature. As a corollary of Theorem 2.14, we have:

- The theories $\mathbf{Q}, I\Sigma_0, I\Sigma_0+\Omega_1, \ldots, I\Sigma_0+\Omega_n, \ldots, B\Sigma_1, B\Sigma_1+\Omega_1, \ldots, B\Sigma_1+\Omega_n, \ldots$ are all mutually interpretable;
- $I\Sigma_0+\mathbf{exp}$ and $B\Sigma_1+\mathbf{exp}$ are mutually interpretable;

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

- For $n \geq 1$, $I\Sigma_n$ and $B\Sigma_{n+1}$ are mutually interpretable;
- $\mathbf{Q} \lhd I\Sigma_0 + \mathbf{exp} \lhd I\Sigma_1 \lhd I\Sigma_2 \lhd \cdots \lhd I\Sigma_n \lhd \cdots \lhd \mathbf{PA}$.

Since any consistent r.e. theory which interprets **Q** is essentially undecidable, G1 holds for all these fragments of **PA** extending **Q**.

Fifth, we discuss some weak theories mutually interpretable with **Q** from the literature. It is interesting to compare **Q** with its bigger brother $\mathbf{PA}^-$. From [132], $\mathbf{PA}^-$ is interpretable in **Q**, and hence **Q** is mutually interpretable with $\mathbf{PA}^-$. The theory $\mathbf{Q}^+$ is interpretable in **Q** (see [35, Theorem 1, p. 296]), and thus mutually interpretable with **Q**. A. Grzegorczyk asks whether $\mathbf{Q}^-$ is essentially undecidable. Švejdar [123] provides a positive answer to Grzegorczyk's original question by showing that **Q** is interpretable in $\mathbf{Q}^-$ using the Solovay's method of shortening cuts. Thus $\mathbf{Q}^-$ is essentially undecidable and mutually interpretable with **Q**.

Sixth, by [35], $I\Sigma_0$ is interpretable in $\mathbf{S}^1_2$, and $\mathbf{S}^1_2$ is interpretable in **Q**. Hence $\mathbf{S}^1_2$ is essentially undecidable and mutually interpretable with **Q**. The theory **AS** interprets Robinson's Arithmetic **Q**, and hence is essentially undecidable. Nelson [98] shows that **AS** is interpretable in **Q**. Thus, **AS** is mutually interpretable with **Q**.

Seventh, Grzegorczyk, and Zdanowski [48] formulate but leave unanswered an interesting problem: are **TC** and **Q** mutually interpretable? M. Ganea [42] proves that **Q** is interpretable in **TC** using the detour via $\mathbf{Q}^-$ (i.e., first show that $\mathbf{Q}^-$ is interpretable in **TC**; since **Q** is interpretable in $\mathbf{Q}^-$, then we have **Q** is interpretable in **TC**). Sterken and Visser [129] give a proof of the interpretability of **Q** in **TC** not using $\mathbf{Q}^-$. Note that **TC** is easily interpretable in the bounded arithmetic $I\Sigma_0$. Thus, **TC** is mutually interpretable with **Q**.

Note that $\mathbf{R} \lhd \mathbf{Q}$ since **Q** is not interpretable in **R** (if **Q** is interpretable in **R**, then **Q** is interpretable in some finite fragment of **R**; however **R** is locally finitely satisfiable and any model of **Q** is infinite). Visser [131] provides us with a unique characterization of **R**.

THEOREM 4.9 (Visser [131, Theorem 6]). *For any consistent r.e. theory T, T is interpretable in* **R** *if and only if T is locally finitely satisfiable.*[23]

Since relational Σ_2 sentences have the finite model property, by Theorem 4.9, any consistent theory axiomatized by a recursive set of Σ_2 sentences in a finite relational language is interpretable in **R**. Since all recursive functions are representable in **R** (see [125, Theorem 6, p. 56]), as a corollary of Theorem 2.6, **R** is essentially undecidable. Cobham shows that **R** has a stronger property than essential undecidability. Vaught gives a proof of Cobham's Theorem 4.10 via existential interpretation in [127].

THEOREM 4.10 (Cobham [127]). *Any consistent r.e. theory that weakly interprets* **R** *is undecidable.*

Eighth, we discuss some variants of **R** in the same language as $L(\mathbf{R}) = \{\overline{0}, \ldots, \overline{n}, \ldots, +, \times, \leq\}$. The theory $\mathbf{R}_0$ is no longer essentially undecidable

[23] In fact, if T is locally finitely satisfiable, then T is interpretable in **R** via a one-piece one-dimensional parameter-free interpretation.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

in the same language as $\mathbf{R}$.[24] In fact, whether $\mathbf{R}_0$ is essentially undecidable depends on the language of $\mathbf{R}_0$: if $L(\mathbf{R}_0) = \{\mathbf{0}, \mathbf{S}, +, \times, \leq\}$ with $\leq$ defined in terms of $+$, then $\mathbf{R}_0$ is essentially undecidable (Cobham first observed that $\mathbf{R}$ is interpretable in $\mathbf{R}_0$ in the same language $\{\mathbf{0}, \mathbf{S}, +, \times\}$, and hence $\mathbf{R}_0$ is essentially undecidable (see [65, 127]). The theory $\mathbf{R}_1$ is essentially undecidable since $\mathbf{R}$ is interpretable in $\mathbf{R}_1$ (see [65, p. 62]).

However $\mathbf{R}_1$ is not minimal essentially undecidable. From [65], $\mathbf{R}$ is interpretable in $\mathbf{R}_2$, and hence $\mathbf{R}_2$ is essentially undecidable.[25] The theory $\mathbf{R}_2$ is minimal essentially undecidable in the sense that if we delete any axiom scheme of $\mathbf{R}_2$, then the remaining system is not essentially undecidable.[26] By essentially the same argument as in [132], we can show that any consistent r.e. theory that weakly interprets $\mathbf{R}_2$ is undecidable.

Kojiro Higuchi and Yoshihiro Horihata introduce the theory of concatenation $\mathbf{WTC}^{-\varepsilon}$, which is a weak subtheory of Grzegorczyk's theory $\mathbf{TC}$, and show that $\mathbf{WTC}^{-\varepsilon}$ is minimal essentially undecidable and $\mathbf{WTC}^{-\varepsilon}$ is mutually interpretable with $\mathbf{R}$ (see [58]).

In summary, we have the following pictures:

- Theories $\mathbf{PA}^-, \mathbf{Q}^+, \mathbf{Q}^-, \mathbf{TC}, \mathbf{AS}, \mathbf{S}^1_2$, and $\mathbf{Q}$ are all mutually interpretable, and hence G1 holds for them;
- Theories $\mathbf{R}, \mathbf{R}_1, \mathbf{R}_2$, and $\mathbf{WTC}^{-\varepsilon}$ are mutually interpretable, and hence G1 holds for them;
- $\mathbf{R} \lhd \mathbf{Q} \lhd \mathbf{EA} \lhd \mathbf{PRA} \lhd \mathbf{PA}$.

4.3.2. The limit of G1 *w.r.t. interpretation and Turing reducibility.* We first discuss the limit of G1 for theories weaker than $\mathbf{PA}$ w.r.t. interpretation, i.e., finding a theory with minimal degree of interpretation for which G1 holds.

First of all, a natural question is: is $\mathbf{Q}$ the weakest finitely axiomatized essentially undecidable theory w.r.t. interpretation such that $\mathbf{R} \lhd \mathbf{Q}$? The following theorem tells us that the answer is no: for any finitely axiomatized subtheory A of $\mathbf{Q}$ that extends $\mathbf{R}$, we can find a finitely axiomatized subtheory B of A such that B extends $\mathbf{R}$ and B does not interpret A.

THEOREM 4.11 (Visser [132, Theorem 2]). *Suppose A is a finitely axiomatized consistent theory and $\mathbf{R} \subseteq A$. Then there is a finitely axiomatized theory B such that $\mathbf{R} \subseteq B \subseteq A$ and $B \lhd A$.*

Define $X = \{S : \mathbf{R} \unlhd S \lhd \mathbf{Q}$ and S is finitely axiomatized$\}$. Theorem 4.11 shows that the structure $\langle X, \lhd \rangle$ is not well-founded.

[24] The theory $\mathbf{R}_0$ has a decidable complete extension given by the theory of reals with $\leq$ as the empty relation on reals.

[25] Another way to show that $\mathbf{R}_2$ is essentially undecidable is to prove that all recursive functions are representable in $\mathbf{R}_2$.

[26] If we delete **Ax2**, then the theory of natural numbers with $x \times y$ defined as $x + y$ is a complete decidable extension; if we delete **Ax3**, then the theory of models with only one element is a complete decidable extension; if we delete **Ax4′**, then the theory of reals is a complete decidable extension.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

THEOREM 4.12 (Visser [132, Theorem 12]). *Suppose A and B are finitely axiomatized theories that weakly interpret* **Q**. *Then there are finitely axiomatized theories $\overline{A} \supseteq A$ and $\overline{B} \supseteq B$ such that $\overline{A}$ and $\overline{B}$ are incomparable (i.e., $\overline{A} \not\trianglelefteq \overline{B}$ and $\overline{B} \not\trianglelefteq \overline{A}$).*

Theorem 4.12 shows that there are incomparable theories extending **Q** w.r.t. interpretation.

Up to now, we do not have an example of essentially undecidable theory that is weaker than **R** w.r.t. interpretation. To this end, we introduce Jeřábek's theory $\mathbf{Rep}_{\mathsf{PRF}}$.

DEFINITION 4.13 (The system $\mathbf{Rep}_{\mathsf{PRF}}$).

- Let PRF denote the sets of all partial recursive functions.
- The language $L(\mathbf{Rep}_{\mathsf{PRF}})$ consists of constant symbols $\overline{n}$ for each $n \in \omega$, and function symbols $\overline{f}$ of appropriate arity for each partial recursive function f.
- The theory $\mathbf{Rep}_{\mathsf{PRF}}$ has axioms:
  - $\overline{n} \neq \overline{m}$ for $n \neq m \in \omega$;
  - $\overline{f}(\overline{n_0}, \ldots, \overline{n_{k-1}}) = \overline{m}$ for each k-ary partial recursive function f such that $f(n_0, \ldots, n_{k-1}) = m$ where $n_0, \ldots, n_{k-1}, m \in \omega$.

The theory $\mathbf{Rep}_{\mathsf{PRF}}$ is essentially undecidable since all recursive functions are representable in it. Since $\mathbf{Rep}_{\mathsf{PRF}}$ is locally finitely satisfiable, by Theorem 4.9, $\mathbf{Rep}_{\mathsf{PRF}} \trianglelefteq \mathbf{R}$. Jeřábek proves that **R** is not interpretable in $\mathbf{Rep}_{\mathsf{PRF}}$. Thus $\mathbf{Rep}_{\mathsf{PRF}} \triangleleft \mathbf{R}$.

Cheng [22] provides more examples of a theory S such that G1 holds for S and $S \triangleleft \mathbf{R}$, and shows that we can find many theories T such that G1 holds for T and $T \triangleleft \mathbf{R}$ based on Jeřábek's work which uses model theory.

THEOREM 4.14 (Cheng [22]). *For any recursively inseparable pair $\langle A, B\rangle$, there is an r.e. theory $U_{\langle A,B\rangle}$ such that* G1 *holds for $U_{\langle A,B\rangle}$, and $U_{\langle A,B\rangle} \triangleleft \mathbf{R}$.*

Define $\mathsf{D} = \{S : S \triangleleft \mathbf{R}$ and G1 holds for theory $S\}$. Theorem 4.14 shows that we could find many witnesses for D. Naturally, we could ask the following questions:

QUESTION 4.15.

- Is $\langle \mathsf{D}, \triangleleft\rangle$ well-founded?
- Are any two elements of $\langle \mathsf{D}, \triangleleft\rangle$ comparable?
- Does there exist a minimal theory w.r.t. interpretation such that G1 holds for it?

We conjectured the following answers to these questions: $\langle \mathsf{D}, \triangleleft\rangle$ is not well founded, $\langle \mathsf{D}, \triangleleft\rangle$ has incomparable elements, and there is no minimal theory w.r.t. interpretation for which G1 holds.

Finally, we discuss the limit of applicability of G1 w.r.t. Turing reducibility. We have discussed the limit of applicability of G1 w.r.t. interpretation. A natural question is: what is the limit of applicability of G1 w.r.t. Turing reducibility?

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

DEFINITION 4.16 (Turing reducibility, the structure $\overline{\mathsf{D}}$).

- Let $\mathcal{R}$ be the structure of the r.e. degrees with the ordering $\leq_T$ induced by Turing reducibility with the least element $\mathbf{0}$ and the greatest element $\mathbf{0}'$.
- Let $\overline{\mathsf{D}} = \{S : S <_T \mathbf{R}$ and G1 holds for theory $S\}$ where $S <_T \mathbf{R}$ stands for $S \leq_T \mathbf{R}$ but $\mathbf{R} \not\leq_T S$.

Cheng [22] shows that for any Turing degree $\mathbf{0} < \mathbf{d} < \mathbf{0}'$, there is a theory U such that G1 holds for U, $U <_T \mathbf{R}$, and U has Turing degree $\mathbf{d}$. As a corollary of this result and known results about the degree structure of $\langle \mathcal{R}, <_T \rangle$ in recursion theory, we can answer above questions for the structure $\langle \overline{\mathsf{D}}, <_T \rangle$:

THEOREM 4.17 (Cheng [22]).

- $\langle \overline{\mathsf{D}}, <_T \rangle$ *is not well-founded*;
- $\langle \overline{\mathsf{D}}, <_T \rangle$ *has incomparable elements*;
- *There is no minimal theory w.r.t. Turing reducibility such that* G1 *holds for it*.

Moreover, Cheng [22] shows that for any Turing degree $\mathbf{0} < \mathbf{d} < \mathbf{0}'$, there is a theory U such that G1 holds for U, $U \trianglelefteq \mathbf{R}$, and U has Turing degree $\mathbf{d}$. Thus, examining the limit of applicability of G1 w.r.t. interpretation is much harder than that w.r.t. Turing reducibility. The structure of $\langle \mathsf{D}, \trianglelefteq \rangle$ is a deep and interesting open question for future research.

§5. The limit of the applicability of G2.

5.1. ***Introduction.*** In our view, G2 is *fundamentally different* from G1. In fact, both mathematically and philosophically, G2 is more problematic than G1 for the following reason. On one hand, in the case of G1, we can construct a natural independent sentence with real mathematical content *without* referring to arithmetization and provability predicates. On the other hand, the meaning of G2 strongly depends on how we exactly formulate the consistency statement.

Similar to [54], we call a result *intensional* if it depends on (the details of) the representation used. Thus, G1 can be called extensional (that is, non-intensional), while G2 is (highly) intensional. We refer to Section 5.3 for more discussion on the intensionality of G2. In this section, we discuss the limit of applicability of G2: under what conditions G2 holds, and under what conditions G2 fails. In Section 5.2, we discuss generalizations of G2.

5.2. ***Some generalizations of*** G2**.** After Gödel, generalizations of G2 are the subject of extensive studies. We know that G2 holds for any consistent r.e. extension of **PA**. However, it is not true that G2 holds for any extension of **PA**. For example, Karl-Georg Niebergall [99] shows that the theory

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

$(\mathbf{PA}+\mathbf{RFN}(\mathbf{PA}))\cap(\mathbf{PA}+$ all true Π^0_1 sentences) can prove its own canonical consistency sentence.[27]

Similarly to G1, one can generalise G2 to arithmetically definable non-r.e. extensions of **PA**. Kikuchi and Kurahashi [71] reformulate G2 as: if S is a Σ^0_1-definable and consistent extension of **PA**, then for any Σ^0_1 definition $\sigma(u)$ of S, $S \nvdash \mathbf{Con}_\sigma(S)$ (see [71, Fact 5.1]). Kikuchi and Kurahashi [71] generalize G2 to arithmetically definable non-r.e. extensions of **PA** and prove that if S is a Σ^0_{n+1}-definable and Σ^0_n-sound extension of **PA**, then there exists a Σ^0_{n+1} definition $\sigma(u)$ of some axiomatization of $Th(S)$ such that $\mathbf{Con}_\sigma(S)$ is independent of S. This corollary shows that the witness for the generalized version of G1 can be provided by the appropriate consistency statement.

Chao–Seraji [20] and Kikuchi–Kurahashi [71] give another generalization of G2 to arithmetically definable non-r.e. extensions of **PA**: for each $n \in \omega$, any Σ^0_{n+1}-definable and Σ^0_n-sound extension of **PA** cannot prove its own Σ^0_n-soundness (see [20, Theorem 2] and [71, Theorem 5.6]). The optimality of this generalization is shown in [20]: there is a Σ^0_{n+1}-definable and Σ^0_{n-1}-sound extension of **PA** that proves its own Σ^0_{n-1}-soundness for $n > 0$ (see [20, Theorem 3]).

Let T be a consistent r.e. extension of **Q**. Kreisel [81] shows that $\neg\mathbf{Con}(T)$ is Π^0_1-conservative over T which is a generalization of G2. We can also generalize G2 via the notion of standard provability predicate.

THEOREM 5.1 *Let T be any consistent r.e. extension of* **Q**. *If $\mathbf{Pr}_T(x)$ is a standard provability predicate, then $T \nvdash \mathbf{Con}(T)$.*

Lev Beklemishev and Daniyar Shamkanov [9] prove that in an abstract setting that presupposes the presence of Gödel's fixed point (instead of directly constructing it, as in the case of formal arithmetic), the Hilbert–Bernays–Löb conditions imply G2 even with fairly minimal conditions on the underlying logic. The following two theorems, due to Feferman and Visser, generalize G2 in terms of the notion of interpretation.

THEOREM 5.2 (Feferman's theorem on the interpretability of inconsistency [32]). *If T is a consistent r.e. extension of* **Q**, *then $T+\neg\mathbf{Con}(T)$ is interpretable in T.*

THEOREM 5.3 (Pudlák [53, 108]). *There is no consistent r.e. theory S such that $(\mathbf{Q}+\mathbf{Con}(S)) \trianglelefteq S$.*[28]

As a corollary of Theorem 5.3, for any consistent r.e. theory S that interprets **Q**, G2 holds for S: $S \nvdash \mathbf{Con}(S)$. The Arithmetic Completeness Theorem tells us that $S \trianglelefteq (\mathbf{Q}+\mathbf{Con}(S))$ (see [130] for the details). As

[27] For the definition of $\mathbf{RFN}(\mathbf{PA})$, we refer to [91]: $\mathbf{RFN}(\mathbf{PA}) = \{\forall x((\Gamma(x) \wedge \mathbf{Pr}_{\mathbf{PA}}(x)) \to \mathbf{Tr}_\Gamma(x)) : \Gamma \text{ arbitrary}\}$.

[28] Instead of Robinson's Arithmetic **Q**, we can as well have taken $\mathbf{S}^1_2$, or $\mathbf{PA}^-$, or $I\Delta_0+\Omega_1$. Moreover, instead of an arithmetical theory we can have employed a string theory like Grzegorczyk's theory **TC** or adjunctive set theory **AS**. All these theories are the same in the sense that they are mutually interpretable (see [133]).

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

a corollary, we have the following version of G2 which highlights the interpretability power of consistency statements.

COROLLARY 5.4 *For any consistent r.e. theory S, we have $S \lhd (\mathbf{Q} + \mathbf{Con}(S))$.*

DEFINITION 5.5 Let T be a consistent extension of $\mathbf{Q}$. A formula $I(x)$ with one free variable (understood as a number variable) is a definable cut in T (in short, a T-cut) if:

- $T \vdash I(\mathbf{0})$;
- $T \vdash \forall x(I(x) \to I(x+1))$;
- $T \vdash \forall x \forall y(y < x \wedge I(x) \to I(y))$.

DEFINITION 5.6 Let $T \supseteq I\Sigma_1$, let J be a T-cut, and let τ be a $\Sigma_0^{\mathbf{exp}}$-definition of T.[29]

- $\mathbf{Pr}_\tau^I(x)$ is the formula $\exists y(I(y) \wedge \mathbf{Proof}_\tau^I(x,y))$ (saying that there is a τ-proof of x in I).
- $\mathbf{Con}_\tau^I$ is the formula $\neg\exists y(I(y) \wedge \mathbf{Proof}_\tau^I(\mathbf{0} \neq \mathbf{0}, y))$.

The following theorem generalizes G2 to definable cuts.

THEOREM 5.7 ([53, Theorem 3.11]). *Let $T \supseteq I\Sigma_1$, let J be a T-cut, and let τ be a $\Sigma_0^{\mathbf{exp}}$-definition of T. Then $T \nvdash \mathbf{Con}_\tau^I$.*

Next, consider a theory U and an interpretation N of the Tarski–Mostowski–Robinson theory $\mathbf{R}$ in U. A U-predicate $\triangle$ is an L-predicate for U, N if it satisfies the following Löb conditions. We write $\triangle A$ for $\triangle(\ulcorner A \urcorner)$, where $\ulcorner A \urcorner$ is the numeral of the Gödel number of A and we interpret the numbers via N. The Gödel numbering is supposed to be fixed and standard.

DEFINITION 5.8 (Löb conditions).

L1: $\vdash A \Rightarrow \vdash \triangle A$.
L2: $\triangle A, \triangle(A \to B) \vdash \triangle B$.
L3: $\triangle A \vdash \triangle\triangle A$.

PROPOSITION 5.9 (Löb's theorem [133, Theorem 3.3.2]). *Suppose that U is a theory, N is an interpretation of the theory $\mathbf{R}$ in U, and $\triangle$ is a U-predicate that is an L-predicate for U, N. Then:*

- *For all U-sentences A we have: if $U \vdash \triangle A \to A$, then $U \vdash A$.*
- *For all U-sentences A we have: $U \vdash \triangle(\triangle A \to A) \to \triangle A$.*

As a corollary of Proposition 5.9, we formulate a general version of G2 which does not mention the notion of provability predicate.

THEOREM 5.10 (Visser [133]). *For all consistent theories U and all interpretations N of $\mathbf{R}$ in U and all L-predicates $\triangle$ for U, N, we have $U \nvdash \neg\triangle\bot$.*

[29] We extend the language $L(\mathbf{PA})$ by a new unary function symbol $\overline{2}^x$ for the xth power of two. The extended language is denoted $L_0(\mathbf{exp})$. A formula is $\Sigma_0^{\mathbf{exp}}$ if it results from atomic formulas of $L_0(\mathbf{exp})$ by iterated application of logical connectives and bounded quantifiers of the form $(\forall x \leq y)$ or $(\exists x \leq y)$ (see [53]).

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

5.3. *The intensionality of* G2. In this section, we discuss the intensionality of G2 which reveals the limit of the applicability of G2.

5.3.1. Introduction. For a consistent theory T, we say that G2 holds for T if the consistency of T is not provable in T. However, this definition is vague, and whether G2 holds for T depends on how we formulate the consistency statement. We refer to this phenomenon as the intensionality of G2. In fact, G2 is essentially different from G1 due to the intensionality of G2: "whether G2 holds for the base theory" depends on how we formulate the consistency statement in the first place.

Both mathematically and philosophically, G2 is more problematic than G1. In the case of G1, we are mainly interested in the fact that *some* sentence is independent of **PA**. We make no claim to the effect that that sentence "really" expresses what we would express by saying "**PA** cannot prove this sentence". But in the case of G2, we are also interested in the content of the consistency statement. We can say that G1 is extensional in the sense that we can construct a concrete independent mathematical statement without referring to arithmetization and provability predicate. However, G2 is intensional and "whether G2 holds for T" depends on varied factors as we will discuss.

In this section, unless stated otherwise, we assume the following:

- T is a consistent r.e. extension of $\mathbf{Q}$;
- the canonical arithmetic formula to express the consistency of the base theory T is $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\mathbf{0} \neq \mathbf{0})$;
- the canonical numbering we use is Gödel's numbering;
- the provability predicate we use is standard;
- the formula representing the set of axioms is Σ^0_1.

The intensionality of Gödel sentence and the consistency statement has been widely discussed from the literature (e.g., Halbach–Visser [54, 55] and Visser [130]). Halbach and Visser examine the sources of intensionality in the construction of self-referential sentences of arithmetic in [54, 55], and argue that corresponding to the three stages of the construction of self-referential sentences of arithmetic, there are at least three sources of intensionality: coding, expressing a property, and self-reference. The three sources of intensionality are not independent of each other, and a choice made at an earlier stage will have influences on the availability of choices at a later stage. Visser [130] locates three sources of indeterminacy in the formalization of a consistency statement for a theory T:

- the choice of a proof system;
- the choice of a way of numbering;
- the choice of a specific formula representing the set of axioms of T.

In summary, the intensional nature ultimately traces back to the various parameter choices that one has to make in arithmetizing the provability predicate. That is the source of both the intensional nature of the Gödel sentence and the consistency sentence.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Based on this and other works from the literature, we argue that "whether G2 holds for the base theory" depends on the following factors:

(1) the choice of the provability predicate (Section 5.3.2);
(2) the choice of the formula expressing consistency (Section 5.3.3);
(3) the choice of the base theory (Section 5.3.4);
(4) the choice of the numbering (Section 5.3.5);
(5) the choice of the formula representing the set of axioms (Section 5.3.6).

These factors are not independent, and a choice made at an earlier stage may have effects on the choices available at a later stage. In the following, unless stated otherwise, when we discuss how G2 depends on one factor, we always assume that other factors are fixed, and only the factor we are discussing is varied. For example, Visser [130] rests on fixed choices for (1)–(2) and (4)–(5) but varies the choice of (3); Grabmayr [46] rests on fixed choices for (1)–(3) and (5) but varies the choice of (4); Feferman [32] rests on fixed choices for (1)–(4) but varies the choice of (5).

5.3.2. The choice of provability predicate. In this section, we show that "whether G2 holds for the base theory" depends on the choice of the provability predicate we use.

As Visser argues in [133], being a consistency statement is not an absolute concept but a role w.r.t. a choice of provability predicate (see Visser [133]). From Theorem 5.1, G2 holds for standard provability predicates. However, G2 may fail for non-standard provability predicates.

Mostowski [96] gives an example of a Σ^0_1 provability predicate for which G2 fails. Let $\mathbf{Pr}^M_T(x)$ be the Σ^0_1 formula "$\exists y(\mathbf{Prf}_T(x,y) \wedge \neg\mathbf{Prf}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner, y))$" where $\mathbf{Prf}_T(x,y)$ is a Δ^0_1 formula saying that "y is a proof of x". Then $\neg\mathbf{Pr}^M_T(\ulcorner 0 \neq 0 \urcorner)$ is trivially provable in **PA**. We know that G2 holds for provability predicates satisfying **D1**– **D3**. Since the formula $\mathbf{Pr}^M_T(x)$ satisfies **D1** and **D3**, it does not satisfy **D2**. Mostowski's example [96] shows that G2 may fail for Σ^0_1 provability predicates satisfying **D1** and **D3**.

One important non-standard provability predicate is Rosser provability predicate $\mathbf{Pr}^R_T(x)$ introduced by Rosser [111] to improve Gödel's first incompleteness theorem. Recall that we have defined the Rosser provability predicate in Definition 3.8. The consistency statement $\mathbf{Con}^R(T)$ based on a Rosser provability predicate $\mathbf{Pr}^R_T(x)$ is naturally defined as $\neg\mathbf{Pr}^R_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$.

It is an easy fact that for any sentence ϕ and Rosser provability predicate $\mathbf{Pr}^R_T(x)$, if $T \vdash \neg\phi$, then $T \vdash \neg\mathbf{Pr}^R_T(\ulcorner \phi \urcorner)$ (see [72, Proposition 2.1]). As a corollary, the consistency statement $\mathbf{Con}^R(T)$ based on Rosser provability predicate $\mathbf{Pr}^R_T(x)$ is provable in T. In this sense, we can say that G2 fails for the consistency statement constructed from Rosser provability predicates.

We can construct different Rosser provability predicates with varied properties. We know that each Rosser provability predicate $\mathbf{Pr}^R_T(x)$ does not satisfy at least one of conditions **D2** and **D3**. Guaspari and Solovay [50] establish a very powerful method of constructing a new proof predicate with required properties from a given proof predicate by reordering nonstandard

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

proofs. Applying this tool, Guaspari and Solovay [50] construct a Rosser provability predicate for which both **D2** and **D3** fail. Arai [2] constructs a Rosser provability predicate with condition **D2**, and a Rosser provability predicate with condition **D3**.

Slow provability, introduced by S. D. Friedman, M. Rathjen, and A. Weiermann [41], is another notion of nonstandard provability for **PA** from the literature. The slow consistency statement $\mathbf{Con}^*(\mathbf{PA})$ asserts that a contradiction is not slow provable in **PA** (for the definition of $\mathbf{Con}^*(\mathbf{PA})$, we refer to [41]). In fact, G2 holds for slow provability: Friedman, Rathjen, and Weiermann show that $\mathbf{PA} \nvdash \mathbf{Con}^*(\mathbf{PA})$ (see [41, Proposition 3.3]). Moreover, Friedman, Rathjen, and Weiermann [41] show that $\mathbf{PA} + \mathbf{Con}^*(\mathbf{PA}) \nvdash \mathbf{Con}(\mathbf{PA})$ (see [41, Theorem 3.10]), and the logical strength of the theory $\mathbf{PA} + \mathbf{Con}^*(\mathbf{PA})$ lies strictly between **PA** and $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$: $\mathbf{PA} \subsetneq \mathbf{PA} + \mathbf{Con}^*(\mathbf{PA}) \subsetneq \mathbf{PA} + \mathbf{Con}(\mathbf{PA})$. Henk and Pakhomov [57] study three variants of slow provability, and show that the associated consistency statement of each of these notions of provability yields a theory that lies strictly between **PA** and $\mathbf{PA} + \mathbf{Con}(\mathbf{PA})$ in terms of logical strength.

5.3.3. The choice of the formula expressing consistency. We show that "whether G2 holds for the base theory" depends on the choice of the arithmetic formula used to express consistency. In the literature, an arithmetic formula is usually used to express the consistency statement. Artemov [3] argues that in Hilbert's consistency program, the original formulation of consistency "no sequence of formulas is a derivation of a contradiction" is about finite sequences of formulas, not about arithmetization, proof codes, and internalized quantifiers.

The canonical consistency statement, the arithmetical formula $\mathbf{Con}(\mathbf{PA})$, says that for all x, x is not a code of a proof of a contradiction in **PA**. In a nonstandard model of **PA**, the universal quantifier "for all x" ranges over both standard and nonstandard numbers, and hence $\mathbf{Con}(\mathbf{PA})$ expresses the consistency of both standard and nonstandard proof codes (see [3]). Thus, $\mathbf{Con}(\mathbf{PA})$ is stronger than the original formulation of consistency which only talks about sequences of formulas and such sequences have only standard codes. Hence, Artemov [3] concludes that G2, saying that **PA** cannot prove $\mathbf{Con}(\mathbf{PA})$, does not actually exclude finitary consistency proofs of the original formulation of consistency (see [3]).

Artemov shows that the original formulation of consistency admits a direct proof in informal arithmetic, and this proof is formalizable in **PA** (see [3]).[30] Artemov's work establishes the consistency of **PA** by finitary means, and vindicates Hilbert's consistency program to some extent.

In the following, we use a single arithmetic sentence to express the consistency statement. Among consistency statements defined via arithmetization,

[30]Informal arithmetic is the theory of informal elementary number theory containing recursive identities of addition and multiplication as well as the induction principle. The formal arithmetic **PA** is just the conventional formalization of the informal arithmetic (see [3]).

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

there are three candidates of arithmetic formulas to express consistency as follows:

- $\mathbf{Con}^0(T) \triangleq \forall x(\mathbf{Fml}(x) \wedge \mathbf{Pr}_T(x) \rightarrow \neg\mathbf{Pr}_T(\dot{\neg}x))$;[31]
- $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$;
- $\mathbf{Con}^1(T) \triangleq \exists x(\mathbf{Fml}(x) \wedge \neg\mathbf{Pr}_T(x))$.

Gödel originally formulates G2 with the consistency statement $\mathbf{Con}^1(T)$: if T is consistent, then $T \nvdash \mathbf{Con}^1(T)$. From the literature, $\mathbf{Con}(T)$ is the widely used canonical consistency statement. Note that $\mathbf{Con}^0(T)$ implies $\mathbf{Con}(T)$, and $\mathbf{Con}(T)$ implies $\mathbf{Con}^1(T)$. However the converse implications do not hold in general (see [88]). Kurahashi [88] proposes different sets of derivability conditions (local version, uniform version, and global version), and examines whether they are sufficient to show the unprovability of these consistency statements (e.g., $\mathbf{Con}(T)$, $\mathbf{Con}^0(T)$, and $\mathbf{Con}^1(T)$).

HB1: If $T \vdash \phi \rightarrow \varphi$, then $T \vdash \mathbf{Pr}_T(\ulcorner \phi \urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner \varphi \urcorner)$.
HB2: $T \vdash \mathbf{Pr}_T(\ulcorner \neg\phi(x) \urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner \neg\phi(\dot{x}) \urcorner)$.
HB3: $T \vdash f(x) = 0 \rightarrow \mathbf{Pr}_T(\ulcorner f(\dot{x}) = 0 \urcorner)$ for every primitive recursive term $f(x)$.

HB1–**HB3** are called the Hilbert–Bernays derivability conditions. If a provability predicate $\mathbf{Pr}_T(x)$ satisfies **HB1**–**HB3**, then $T \nvdash \mathbf{Con}^0(T)$ (see [59]). Kurahashi [89] constructs two Rosser provability predicates satisfying **HB1**–**HB3**. Thus, **HB1**–**HB3** are not sufficient to prove that $T \nvdash \mathbf{Con}(T)$.

Löb [92] proves that if $\mathbf{Pr}_T(x)$ satisfies the Hilbert–Bernays–Löb derivability conditions **D1**–**D3** (see Definition 3.6), then Löb's theorem holds: for any sentence ϕ, if $T \vdash \mathbf{Pr}_T(\ulcorner \phi \urcorner) \rightarrow \phi$, then $T \vdash \phi$. It is well-known that Löb's theorem implies G2: $T \nvdash \mathbf{Con}(T)$ (see [97]). Thus, if a provability predicate $\mathbf{Pr}_T(x)$ satisfies **D1**–**D3**, then $T \nvdash \mathbf{Con}(T)$. Kurahashi [88, Proposition 4.11] constructs a provability predicate $\mathbf{Pr}_T(x)$ with conditions **D1**–**D3**, but $T \vdash \mathbf{Con}^1(T)$. Thus, **D1**–**D3** are not sufficient to prove that $T \nvdash \mathbf{Con}^1(T)$.

Montagna [94] proves that if a provability predicate $\mathbf{Pr}_T(x)$ satisfies the following two conditions, then $T \nvdash \mathbf{Con}^1(T)$:

- $T \vdash \forall x(\text{“x is a logical axiom”} \rightarrow \mathbf{Pr}_T(x))$;
- $T \vdash \forall x \forall y(\mathbf{Fml}(x) \wedge \mathbf{Fml}(y) \rightarrow (\mathbf{Pr}_T(x \rightarrow y) \rightarrow (\mathbf{Pr}_T(x) \rightarrow \mathbf{Pr}_T(y))))$.

5.3.4. The choice of base theory. We show that “whether G2 holds for the base theory” depends on the base theory we choose. A foundational question about G2 is: how much information about arithmetic is required for the proof of G2? If the base system does not contain enough information about arithmetic, then G2 may fail. The widely used notion of consistency is consistency in proof systems with cut elimination. However, notions like cutfree consistency, Herbrand consistency, tableaux consistency, and restricted consistency for different base theories behave differently (see [130]). We do have proof systems that prove their own cutfree consistency:

[31] $\mathbf{Fml}(x)$ is the formula which represents the relation that x is a code of a formula.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

for example, finitely axiomatized sequential theories prove their own cut-free consistency on a definable cut (see [136, p. 25]).

A natural question is: whether G2 can be generalized to base systems weaker than **PA** w.r.t. interpretation? As a corollary of Theorem 5.3, we have $\mathbf{Q} \nvdash \mathbf{Con}(\mathbf{Q})$ and hence $\mathbf{Q} \nvdash \mathbf{Con}^0(\mathbf{Q})$. Bezboruah and Shepherdson [11] define the consistency of **Q** as the sentence $\mathbf{Con}^0(\mathbf{Q})$[32], and prove that G2 holds for $\mathbf{Q} : \mathbf{Q} \nvdash \mathbf{Con}^0(\mathbf{Q})$. However, the method used by Bezboruah and Shepherdson in [11] is quite different from Theorem 5.3. Bezboruah–Shepherdson's proof depends on some specific assumptions about the coding, does not easily generalize to stronger theories, and tells us nothing about the question whether **Q** can prove its consistency on some definable cut (see [132]). The next question is: whether G2 holds for other theories weaker than **Q** w.r.t. interpretation (e.g., **R**)? In a forthcoming paper, we will show that G2 holds for **R** via the canonical consistency statement. However, we can find weak theories mutually interpretable with **R** for which G2 fails.

Willard [145] explores the generality and boundary-case exceptions of G2 over some base theories. Willard constructs examples of r.e. arithmetical theories that cannot prove the totality of their successor functions but can prove their own canonical consistencies (see [144, 145]). However, the theories Willard constructs are not completely natural since some axioms are constructed using Gödel's Diagnolisation Lemma. Pakhomov [102] constructs a more natural example of this kind. Pakhomov [102] defines a theory $H_{<\omega}$, and shows that it proves its own canonical consistency. Unlike Willard's theories, $H_{<\omega}$ isn't an arithmetical theory but a theory formulated in the language of set theory with an additional unary function. From [102], $H_{<\omega}$ and **R** are mutually interpretable. Hence, the theory $H_{<\omega}$ can be regarded as the set-theoretic analogue of **R** from the interpretability theoretic point of view.

From Theorem 5.3, G2 holds for any consistent r.e. theory interpreting **Q**. However, it is not true that G2 holds for any consistent r.e. theory interpreting **R** since $H_{<\omega}$ interprets **R**, but G2 fails for $H_{<\omega}$. We know that if $S \trianglelefteq T$ and G1 holds for S, then G1 holds for T. However, it is not true that if $S \trianglelefteq T$ and G2 holds for S, then G2 holds for T since $\mathbf{R} \trianglelefteq H_{<\omega}$, G2 holds for **R** but G2 fails for $H_{<\omega}$. This shows the difference between **Q** and **R**, and the difference between G1 and G2.

One way to eliminate the intensionality of G2 is to uniquely characterize the consistency statement. In [130], Visser proposes the interesting question of a coordinate-free formulation of G2 and a unique characterization of the consistency statement. Visser [130] shows that consistency for finitely axiomatized sequential theories can be uniquely characterized modulo **EA**-provable equivalence (see [130, p. 543]). But characterizing the consistency of infinitely axiomatized r.e. theories is more delicate and a big open problem in the current research on the intensionality of G2.

[32]This sentence says that for any x, if x is the code of a formula ϕ and ϕ is provable in **Q**, then $\neg\phi$ is not provable in **Q**.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

After Gödel, Gentzen constructs a theory $\mathbf{T}^*$ (primitive recursive arithmetic with the additional principle of quantifier-free transfinite induction up to the ordinal ε_0),[33] and proves the consistency of **PA** in $\mathbf{T}^*$. Gentzen's theory $\mathbf{T}^*$ contains **Q** but does not contain **PA** since $\mathbf{T}^*$ does not prove the ordinary mathematical induction for all formulas. By the Arithmetized Completeness Theorem, $\mathbf{Q}+\mathbf{Con}(\mathbf{PA})$ interprets **PA**. Since Gentzen's theory $\mathbf{T}^*$ contains **Q** and $\mathbf{T}^* \vdash \mathbf{Con}(\mathbf{PA})$, Gentzen's theory $\mathbf{T}^*$ interprets **PA**. By Pudlák's result that no consistent r.e. extension T of **Q** can interpret $\mathbf{Q}+\mathbf{Con}(T)$, **PA** does not interpret Gentzen's theory $\mathbf{T}^*$. Thus $\mathbf{PA} \lhd \mathbf{T}^*$. Gentzen's work has opened a productive new direction in proof theory: finding the means necessary to prove the consistency of a given theory. More powerful subsystems of Second-Order Arithmetic have been given consistency proofs by Gaisi Takeuti and others, and theories that have been proved consistent by these methods are quite strong and include most ordinary mathematics.

5.3.5. The choice of numbering. We show that "whether G2 holds for the base theory" depends on the choice of the numbering encoding the language.

For the influence of different numberings on G2, we refer to [46]. Any injective function γ from a set of $L(\mathbf{PA})$-expressions to ω qualifies as a numbering. Gödel's numbering is a special kind of numberings under which the Gödel number of the set of axioms of **PA** is recursive. In fact, G2 is sensitive to the way of numberings. Let γ be a numbering and $\ulcorner\varphi^{\gamma}\urcorner$ denote $\overline{\gamma(\varphi)}$, i.e., the standard numeral of the γ-code of φ.

Definition 5.11 (Relativized Löb conditions). A formula $\mathbf{Pr}^{\gamma}_{T}(x)$ is said to satisfy Löb's conditions relative to γ for the base theory T if for all $L(\mathbf{PA})$-sentences φ and ψ we have that:

D1*: If $T \vdash \varphi$, then $\mathbf{PA} \vdash \mathbf{Pr}^{\gamma}_{T}(\ulcorner\varphi^{\gamma}\urcorner)$;
D2*: $T \vdash \mathbf{Pr}^{\gamma}_{T}(\ulcorner(\varphi \to \psi)^{\gamma}\urcorner) \to (\mathbf{Pr}^{\gamma}_{T}(\ulcorner\varphi^{\gamma}\urcorner) \to \mathbf{Pr}^{\gamma}_{T}(\ulcorner\psi^{\gamma}\urcorner))$;
D3*: $T \vdash \mathbf{Pr}^{\gamma}_{T}(\ulcorner\varphi^{\gamma}\urcorner) \to \mathbf{Pr}^{\gamma}_{T}(\ulcorner(\mathbf{Pr}^{\gamma}_{T}(\ulcorner\varphi^{\gamma}\urcorner))^{\gamma}\urcorner)$.

Grabmayr [46] examines different criteria of acceptability, and proves the invariance of G2 with regard to acceptable numberings (for the definition of acceptable numberings, we refer to [46]).

Theorem 5.12 (Invariance of G2 under acceptable numberings [46, Theorem 4.8]). *Let γ be an acceptable numbering and T be a consistent r.e. extension of* **Q**. *If $\mathbf{Pr}^{\gamma}_{T}(x)$ satisfies Löb's conditions* **D1***–**D3*** *relative to γ for T, then $T \nvdash \neg\mathbf{Pr}^{\gamma}_{T}(\ulcorner(\mathbf{0} \neq \mathbf{0})^{\gamma}\urcorner)$.*

Theorem 5.12 shows that G2 holds for acceptable numberings. But G2 may fail for non-acceptable numberings. Grabmayr [46] gives some examples of deviant numberings γ such that G2 fails w.r.t. γ: $T \vdash \mathbf{Pr}^{\gamma}_{T}(\ulcorner(\mathbf{0} \neq \mathbf{0})^{\gamma}\urcorner)$.

Definition 5.13 We say that $\alpha(x)$ is a numeration of T if for any n, we have $\mathbf{PA} \vdash \alpha(\overline{n})$ if and only if n is the Gödel number of some $\phi \in T$.

[33] ε_0 is the first ordinal α such that $\omega^{\alpha} = \alpha$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

5.3.6. The choice of the formula representing the set of axioms. We show that "whether G2 holds for T" depends on the way the axioms of T are represented.

First of all, Definition 5.14 gives a more general definition of provability predicate and consistency statement for T w.r.t. the numeration of T.

DEFINITION 5.14 Let T be any consistent r.e. extension of $\mathbf{Q}$ and $\alpha(x)$ be a formula in $L(T)$.

- Define the formula $\mathbf{Prf}_\alpha(x,y)$ saying "y is the Gödel number of a proof of the formula with Gödel number x from the set of all sentences satisfying $\alpha(x)$".
- Define the provability predicate $\mathbf{Pr}_\alpha(x)$ of $\alpha(x)$ as $\exists y \mathbf{Prf}_\alpha(x,y)$ and the consistency statement $\mathbf{Con}_\alpha(T)$ as $\neg\mathbf{Pr}_\alpha(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$.

For each formula $\alpha(x)$, we have:

$$\mathbf{D2'} \quad \mathbf{PA} \vdash \mathbf{Pr}_\alpha(\ulcorner \varphi \to \psi \urcorner) \to (\mathbf{Pr}_\alpha(\ulcorner \varphi \urcorner) \to \mathbf{Pr}_\alpha(\ulcorner \psi \urcorner)).$$

If $\alpha(x)$ is a numeration of T, then $\mathbf{Pr}_\alpha(x)$ satisfies the following properties (see [86, Fact 2.2]):

D1′: If $T \vdash \varphi$, then $\mathbf{PA} \vdash \mathbf{Pr}_\alpha(\ulcorner \varphi \urcorner)$;
D3′: If φ is Σ^0_1, then $\mathbf{PA} \vdash \varphi \to \mathbf{Pr}_\alpha(\ulcorner \varphi \urcorner)$.

Now we give a new reformulation of G2 via numerations.

THEOREM 5.15 *Let T be any consistent r.e. extension of $\mathbf{Q}$. If $\alpha(x)$ is any Σ^0_1 numeration of T, then $T \nvdash \mathbf{Con}_\alpha(T)$.*

In fact, G2 holds for any Σ^0_1 numeration of T, but fails for some Π^0_1 numeration of T. Feferman [32] constructs a Π^0_1 numeration $\pi(x)$ of T such that G2 fails, i.e., $\mathbf{Con}_\pi(T) \triangleq \neg\mathbf{Pr}_\pi(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$ is provable in T. Feferman's construction keeps the proof predicate and its numbering fixed but varies the formula representing the set of axioms. Notice that Feferman's predicate satisfies **D1** and **D2**, but does not satisfy **D3**. Feferman's example shows that G2 may fail for provability predicates satisfying **D1** and **D2**.

Generally, Feferman [32] shows that if T is a Σ^0_1-definable extension of $\mathbf{Q}$, then there is a Π^0_1 definition $\tau(u)$ of T such that $T \vdash \mathbf{Con}_\tau(T)$. In summary, G2 is not coordinate-free (it is dependent on numerations of **PA**). An important question is how to formulate G2 in a general way such that it is coordinate-free (independent of numerations of T).

The properties of the provability predicate are intensional and depend on the numeration of the theory. I.e., under different numerations of T, the provability predicate may have different properties. It may happen that T has two numerations $\alpha(x)$ and $\beta(x)$ such that $\mathbf{Con}_\alpha(T)$ is not equivalent to $\mathbf{Con}_\beta(T)$. For example, under Gödel's recursive numeration $\tau(x)$ and Feferman's Π^0_1 numeration $\pi(x)$ of T, the corresponding consistency statements $\mathbf{Con}_\tau(T)$ and $\mathbf{Con}_\pi(T)$ are not equivalent. But **PA** does not know this fact, i.e., $\mathbf{PA} \nvdash \neg(\mathbf{Con}_\tau(T) \leftrightarrow \mathbf{Con}_\pi(T))$ since $\mathbf{PA} \nvdash \neg\mathbf{Con}_\tau(T)$.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Generally, Kikuchi and Kurahashi prove in [71, Corollary 5.11] that if T is Σ^0_{n+1}-definable and not Σ^0_n-sound, then there are Σ^0_{n+1} definitions $\sigma_1(x)$ and $\sigma_2(x)$ of T such that $T \vdash \mathbf{Con}_{\sigma_1}(T)$ and $T \vdash \neg\mathbf{Con}_{\sigma_2}(T)$.

Provability logic is an important tool for the study of incompleteness and meta-mathematics of arithmetic. The origins of provability logic (e.g., Henkin's problem, the isolation of derivability conditions, and Löb's theorem) are all closely tied to Gödel's incompleteness theorems historically. In this sense, we can say that Gödel's incompleteness theorems play a unifying role between first order arithmetic and provability logic.

Provability logic is the logic of properties of provability predicates. Note that G2 is very sensitive to the properties of the provability predicate used in its formulation. Provability logic provides us with a new viewpoint and an important tool that can be used to understand incompleteness. Provability logic based on different provability predicates reveals the intensionality of provability predicates which is one source of the intensionality of G2.

Let T be a consistent r.e. extension of $\mathbf{Q}$, and $\tau(u)$ be any numeration of T. Recall that an arithmetical interpretation f is a mapping from the set of all modal propositional variables to the set of $L(T)$-sentences. Every arithmetical interpretation f is uniquely extended to the mapping f_τ from the set of all modal formulas to the set of $L(T)$-sentences such that f_τ satisfies the following conditions:

- $f_\tau(p)$ is $f(p)$ for each propositional variable p;
- $f_\tau(\bot)$ is $\mathbf{0} \neq \mathbf{0}$;
- f_τ commutes with every propositional connective;
- $f_\tau(\Box A)$ is $\mathbf{Pr}_\tau(\ulcorner f_\tau(A) \urcorner)$ for every modal formula A.

Provability logic provides us with a new way of examining the intensionality of provability predicates. Under different numerations of T, the provability predicate may have different properties, and hence may correspond to different modal principles under different arithmetical interpretations.

Definition 5.16 Given a numeration $\tau(u)$ of T, the provability logic $\mathbf{PL}_\tau(T)$ of $\tau(u)$ is defined to be the set of modal formulas A such that $T \vdash f_\tau(A)$ for all arithmetical interpretations f.

Note that the provability logic $\mathbf{PL}_\tau(T)$ of a Σ^0_n numeration $\tau(x)$ of T is a normal modal logic. A natural and interesting question is: which normal modal logic can be realized as a provability logic $\mathbf{PL}_\tau(T)$ of some Σ^0_n numeration $\tau(x)$ of T? An interesting research program is to classify the provability logic $\mathbf{PL}_\alpha(T)$ according to the numeration $\alpha(x)$ of T. We first discuss Σ^0_1 numerations of T.

Theorem 5.17 (Generalized Solovay's Arithmetical Completeness Theorem [86, Theorem 2.5]). *Let T be any consistent r.e. extension of* **PA**. *If T is Σ^0_1-sound, then for any Σ^0_1 numeration $\alpha(x)$ of T, the provability logic $\mathbf{PL}_\alpha(T)$ is precisely* **GL**.[34]

[34] It is a big open problem that whether Solovay's arithmetical completeness theorem holds for weak arithmetic.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

Moreover, Visser [128] examines all possible provability logics for Σ^0_1 numerations of Σ^0_1-unsound theories. To state Visser's result, we need some definitions.

DEFINITION 5.18 ([87, Definitions 3.5 and 3.6]). We define the sequence $\{\mathbf{Con}^n_\tau : n \in \omega\}$ recursively as follows: $\mathbf{Con}^0_\tau$ is $\mathbf{0} = \mathbf{0}$, and $\mathbf{Con}^{n+1}_\tau$ is $\neg\mathbf{Pr}_\tau(\ulcorner\neg\mathbf{Con}^n_\tau\urcorner)$. The height of $\tau(u)$ is the least natural number n such that $T \vdash \neg\mathbf{Con}^n_\tau$ if such an n exists. If not, the height of $\tau(u)$ is ∞.

For Σ^0_1-unsound theories, Visser proves that $\mathbf{PL}_\tau(T)$ is determined by the height of the numeration $\tau(u)$. Visser [128, Theorem 3.7] shows that the height of $\tau(u)$ is ∞ if and only if $\mathbf{PL}(\tau) = \mathbf{GL}$; and the height of $\tau(u)$ is n if and only if $\mathbf{PL}(\tau) = \mathbf{GL} + \Box^n\bot$. Beklemishev [6, Lemma 7] shows that if T is Σ^0_1-unsound, then the height of Σ^0_1 numerations of T can take any values except 0.

Let U be any consistent theory of arithmetic. Based on the previous work by Artemov, Visser, and Japaridze, Beklemishev [6] proves that for Σ^0_1 numeration τ of U, $\mathbf{PL}_\tau(U)$ coincides with one of the logics $\mathbf{GL}_\alpha, \mathbf{D}_\beta, \mathbf{S}_\beta$, and $\mathbf{GL}^-_\beta$ where α and β are subsets of ω and β is cofinite (for definitions of $\mathbf{GL}_\alpha, \mathbf{D}_\beta, \mathbf{S}_\beta$ and $\mathbf{GL}^-_\beta$, we refer to [6]).

Feferman [32] constructs a Π^0_1 numeration $\pi(x)$ of T such that the consistency statement $\mathbf{Con}_\pi(T)$ defined via $\mathbf{Pr}_\pi(x)$ is provable in T. Thus, the provability logic $\mathbf{PL}_\pi(T)$ of $\mathbf{Pr}_\pi(x)$ contains the formula $\neg\Box\bot$, and is different from $\mathbf{GL}$. However, the exact axiomatization of the provability logic $\mathbf{PL}_\pi(T)$ under Feferman's numeration $\pi(x)$ is not known. Kurahashi [86] proves that for any recursively axiomatized consistent extension T of $\mathbf{PA}$, there exists a Σ^0_2 numeration $\alpha(x)$ of T such that the provability logic $\mathbf{PL}_\alpha(T)$ is the modal system $\mathbf{K}$. As a corollary, the modal principles commonly contained in every provability logic $\mathbf{PL}_\alpha(T)$ of T is just $\mathbf{K}$.

It is often thought that a provability predicate satisfies **D1**–**D3** if and only if G2 holds (i.e., for the induced consistency statement $\mathbf{Con}(T)$ from the provability predicate, $T \nvdash \mathbf{Con}(T)$). But this is not true. From Definition 5.14, conditions **D1** and **D2** hold for any numeration of T. Whether the provability predicate satisfies condition **D3** depends on the numeration of T. For any Σ^0_1-numeration $\alpha(x)$ of T, **D3** holds for $\mathbf{Pr}_\alpha(x)$. From Kurahashi [86], there is a Σ^0_2-numeration $\alpha(x)$ of T such that the provability logic for that numeration is precisely $\mathbf{K}$. Since $\mathbf{K} \nvdash \neg\Box\bot$, as a corollary, G2 holds for T, i.e., $\mathbf{Con}_\alpha(T)$ defined as $\neg\mathbf{Pr}_\alpha(\ulcorner\mathbf{0} \neq \mathbf{0}\urcorner)$ is not provable in T. But the Löb condition **D3** does not hold since $\mathbf{K} \nvdash \Box A \to \Box\Box A$. This gives us an example of a Σ^0_2 numeration $\alpha(x)$ of T such that **D3** does not hold for $\mathbf{Pr}_\alpha(x)$ but G2 holds for T. Thus, G2 may hold for a provability predicate which does not satisfy the Löb condition **D3**.

Moreover, Kurahashi [87] proves that for each $n \geq 2$, there exists a Σ^0_2 numeration $\tau(x)$ of T such that the provability logic $\mathbf{PL}_\tau(T)$ is just the modal logic $\mathbf{K} + \Box(\Box^n p \to p) \to \Box p$. Hence there are infinitely many normal modal logics that are provability logics for some Σ^0_2 numeration of T. A good question from Kurahashi [87] is: for $n \geq 2$, is the class of provability logics

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

$\mathbf{PL}_\tau(T)$ for Σ^0_n numerations $\tau(x)$ of T the same as the class of provability logics $\mathbf{PL}_\tau(T)$ for Σ^0_{n+1} numerations $\tau(x)$ of T? However, this question is still open as far as we know. Define that $\mathbf{KD} = \mathbf{K} + \neg\Box\bot$. A natural and interesting question, which is also open as far as we know, is: can we find a numeration $\tau(x)$ of T such that $\mathbf{PL}_\tau(T) = \mathbf{KD}$?

In summary, G2 is intensional with respect to the following parameters: the formalization of consistency, the base theory, the method of numbering, the choice of a provability predicate, and the representation of the set of axioms. Current research on incompleteness reveals that G2 is a deep and profound theorem both mathematically and philosophically in the foundations of mathematics, and there is a lot more to be explored about the intensionality of G2.

§6. Conclusion. We conclude this paper with some personal comments. To the author, the research on concrete incompleteness is very deep and important.

After Gödel, people have found many different proofs of incompleteness theorems via pure logic, and many concrete independent statements with real mathematical contents. As Harvey Friedman comments, the research on concrete mathematical incompleteness shows how the Incompleteness Phenomena touch normal concrete mathematics, and reveal the impact and significance of the foundations of mathematics.

Harvey Friedman's research project on concrete incompleteness plans to show that we will be able to find, in just about any subject of mathematics, many natural looking statements that are independent of **ZFC**. Harvey Friedman's work is very profound and promising, and will reveal that incompleteness is everywhere in mathematics, which, if it is true, may be one of the most important discoveries after Gödel in the foundations of mathematics.

Acknowledgment. I would like to thank Matthias Baaz, Ulrich Kohlenbach, Taishi Kurahashi, Zachiri McKenzie, Fedor Pakhomov, Michael Rathjen, Saeed Salehi, Sam Sanders, and Albert Visser for their valuable comments on this work. Especially, I would like to thank Sam Sanders and Zachiri McKenzie for their proofreading of my English writing of this paper. I would like to thank the referees for providing detailed and helpful comments for improvements. I thank the lamp for my feet, and the light on my path. This paper is the research result of the Humanities and Social Sciences of Ministry of Education Planning Fund project "Research on Gödel's incompleteness theorem" (project no: 17YJA72040001). I would like to thank the fund support by the Humanities and Social Sciences of Ministry of Education Planning Fund.

Some materials of my old paper "Note on some misinterpretations of Gödel's incompleteness theorems" have been incorporated into this paper.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

REFERENCES

[1] Z. Adamowicz and T. Bigorajska, *Existentially closed structures and Gödel's second incompleteness theorem*. ***The Journal of Symbolic Logic***, vol. 66 (2001), no. 1, pp. 349–356.

[2] T. Arai, *Derivability conditions on Rosser's provability predicates*. ***Notre Dame Journal of Formal Logic***, vol. 31 (1990), no. 4, pp. 487–497.

[3] S. N. Artemov, *The provability of consistency*, preprint, 2019, arXiv:1902.07404v5.

[4] J. Avigad, *Incompleteness via the halting problem*, 2005, https://www.andrew.cmu.edu/user/avigad/Teaching/halting.pdf.

[5] K. J. Barwise, *Comments introducing Boolos' article*. ***Notices of the American Mathematical Society***, vol. 36 (1989), no. 388.

[6] L. D. Beklemishev, *On the classification of propositional provability logics*. ***Izvestiya Akademii Nauk SSSR, Seriya Matematicheskaya***, vol. 53 (1989), no. 5, pp. 915–943, translated in ***Mathematics of the USSR-Izvestiya***, vol. 35 (1990), no. 2, pp. 247–275.

[7] ———, ***The Worm Principle***, Logic Group Preprint Series, vol. 219, Utrecht University, the Netherlands, 2003.

[8] ———, *Gödel incompleteness theorems and the limits of their applicability I*. *Russian Mathematical Surveys*, vol. 65 (2010), no. 5, pp. 857–899.

[9] L. D. Beklemishev and D. S. Shamkanov, *Some abstract versions of Gödel's second incompleteness theorem based on non-classical logics*, ***A Tribute to Albert Visser*** (L. A. Alberti, editor), College Publications, 2016, pp. 15–29.

[10] C. Berline, K. Mcaloon, and J. P. Ressayre (eds.), ***Model Theory and Arithmetic***, Lecture Notes in Mathematics, vol. 890, Springer, Berlin, 1981.

[11] A. Bezboruah and J. C. Shepherdson, *Gödel's second incompleteness theorem for* $\boldsymbol{Q}$. ***The Journal of Symbolic Logic***, vol. 41 (1976), no. 2, pp. 503–512.

[12] R. Blanck, ***Contributions to the Metamathematics of Arithmetic: Fixed Points, Independence, and Flexibility***, Ph.D. thesis, University of Gothenburg, Acta Universitatis Gothoburgensis, 2017.

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

[14] ———, ***The Logic of Provability***, Cambridge University Press, Cambridge, 1995.

[15] A. Bovykin, *Brief introduction to unprovability*. ***Logic Colloquium***, Lecture Notes in Logic, vol. 32, Cambridge University Press, Cambridge, 2009, pp 38–64.

[16] B. Buldt, *The scope of Gödel's first incompleteness theorem*. ***Logica Universalis***, vol. 8 (2014), no. (3–4), pp. 499–552.

[17] S. R. Buss, ***Bounded Arithmetic***, Studies in Proof Theory, Lecture Notes, vol. 3, Bibliopolis, Naples, 1986.

[18] ———, *First-order theory of arithmetic*. ***Handbook Proof Theory***, Elsevier, Amsterdam, 1998, pp 79–148.

[19] G. J. Chaitin, *Information-theoretic limitations of formal systems*. ***Journal of the Association for Computing Machinery***, vol. 21 (1974), pp. 403–424.

[20] C. Chao and P. Seraji, *Gödel's second incompleteness theorem for* Σ_n*-definable theories*. ***Logic Journal of the IGPL***, vol. 26 (2018), no. 2, pp. 255–257.

[21] Y. Cheng, *Incompleteness for Higher-Order Arithmetic*: *An Example Based on Harrington's Principle*, Springer Series: Springer, Briefs in Mathematics, Springer, 2019.

[22] ———, *Finding the limit of incompleteness I*, this Journal, 2020, to appear, doi:10.1017/bsl.2020.09.

[23] ———, *On the depth of Gödel's incompleteness theorem*, preprint, 2020, arXiv:2008.13142.

[24] Y. Cheng and R. Schindler, *Harrington's principle in higher order arithmetic*. ***The Journal of Symbolic Logic***, vol. 80, (2015), no. 02, pp 477–489.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

[25] C. Cieśliński, *Heterologicality and incompleteness*. ***Mathematical Logic Quarterly***, vol. 48 (2002), no. 1, pp. 105–110.

[26] C. Cieśliński and R. Urbaniak, *Gödelizing the Yablo sequence*. ***Journal of Philosophical Logic***, vol. 42 (2013), no. (5), pp. 679–695.

[27] P. Clote and K. Mcaloon, *Two further combinatorial theorems equivalent to the* 1*-consistency of Peano arithmetic*. ***Journal of Symbolic Logic***, vol. 48 (1983), no.4, pp. 1090–1104.

[28] W. Dean, *Incompleteness via paradox and completeness*. ***Review of Symbolic Logic***, vol. 13 (2020), no. 3, pp. 541–592.

[29] H. B. Enderton, ***A Mathematical Introduction to Logic***, second ed., Academic Press, Boston, MA, 2001.

[30] ———, ***Computability Theory, An Introduction to Recursion Theory***, Academic Press, Cambridge, MA, 2011.

[31] H. B. Enderton (with contributions by L. W. Szczerba), ***Classical Mathematical Logic: The Semantic Foundations of Logic***, Princeton University Press, 2011.

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

[33] ———, *Transfinite recursive progressions of axiomatic theories*. ***Journal of Symbolic Logic***, vol. 27 (1962), pp. 259–316.

[34] ———, *The impact of the incompleteness theorems on mathematics*. ***Notices of the AMS***, vol. 53 (2006), no. 4, pp. 434–439.

[35] F. Ferreira and G. Ferreira, *Interpretability in Robinson's Q*. this Journal, vol. 19 (2013), no. 3, pp. 289–317.

[36] F. B. Fitch, *A Gödelized formulation of the prediction paradox*. ***American Philosophical Quarterly***, vol. 1 (1964), pp. 161–164.

[37] T. Franzen, ***Inexhaustibility: A Non-Exhaustive Treatment***, Lecture Notes in Logic, vol. 16, Cambridge University Press, Cambridge, 2017.

[38] H. Friedman, *On the necessary use of abstract set theory*. ***Advances in Mathematics***, vol. 41 (1981), pp. 209–280.

[39] ———, *Finite functions and the necessary use of large cardinals*. ***Annals of Mathematics***, vol. 148 (1998), pp. 803–893.

[40] ———, *Boolean Relation Theory and Incompleteness*, Manuscript, to appear.

[41] S.-D. Friedman, M. Rathjen, and A. Weiermann, *Slow consistency*. ***Annals of Pure and Applied Logic***, vol. 164 (2013), no. 3, pp. 382–393.

[42] M. Ganea, *Arithmetic on semigroups*. ***Journal of Symbolic Logic***, vol. 74 (2009), no. 1, pp. 265–278.

[43] K. Gödel, *Über formal unentscheidbare sätze der Principia Mathematica und verwandter systeme I*. ***Monatshefte für Mathematik und Physik***, vol. 38 (1931), no. 1, pp. 173–198.

[44] ———, ***Kurt Gödel's Collected Works, vol. 1: Publications 1929–1936***, Oxford University Press, New York, 1986, pp. 145–195.

[45] L. Gordeev and A. Weiermann, *Phase transitions of iterated Higman-style well-partial-orderings*. ***Archive for Mathematical Logic***, vol. 51 (2012), no. 1–2, pp. 127–161.

[46] B. Grabmayr, *On the invariance of Gödel's second theorem with regard to numberings*. ***The Review of Symbolic Logic***, 2020, doi: 10.1017/S1755020320000192.

[47] A. Grzegorczyk, *Undecidability without arithmetization*. ***Studia Logica***, vol. 79 (2005), no. 2, pp. 163–230.

[48] A. Grzegorczyk and K. Zdanowski, *Undecidability and concatenation*, ***Andrzej Mostowski and Foundational Studies*** (A. Ehrenfeucht, V. W. Marek, and M. Srebrny, editors), IOS Press, Amsterdam, 2008, pp. 72–91.

[49] D. Guaspari, *Partially conservative extensions of arithmetic*. ***Transactions of the American Mathematical Society***, vol. 254 (1979), pp. 47–68.

[50] D. Guaspari and R. M. Solovay, *Rosser sentences*. ***Annals of Mathematical Logic***, vol. 16 (1979), no. 1, pp. 81–99.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

[51] P. Hájek, *Interpretability and fragments of arithmetic*, ***Arithmetic, Proof Theory and Computational Complexity*** (P. Clote and J. Krajícek, editors), Oxford Logic Guides, 23, Clarendon Press, Oxford, 1993, pp. 185–196.

[52] P. Hájek and J. Paris, *Combinatorial principles concerning approximations of functions*. ***Archive for Mathematical Logic***, vol. 26 (1986), no. 1–2, pp. 13–28.

[53] P. Hájek and P. Pudlák, ***Metamathematics of First-Order Arithmetic***, Perspectives in Logic, vol. 3, Cambridge University Press, Cambridge, 2017.

[54] V. Halbach and A. Visser, *Self-reference in arithmetic I*. ***Review of Symbolic Logic***, vol. 7 (2014a), no. 4, pp. 671–691.

[55] ———, *Self-reference in arithmetic II*. ***Review of Symbolic Logic***, vol. 7 (2014b), no. (4), pp. 692–712.

[56] M. Hamano and M. Okada, *A relationship among Gentzen's proof-reduction, Kirby–Paris' hydra game, and Buchholz's hydra game*. ***Mathematical Logic Quarterly***, vol. 43 (1997), no. 1, pp. 103–120.

[57] P. Henk and F. Pakhomov, *Slow and ordinary provability for Peano arithmetic*, preprint, 2016, arXiv:1602.1822.

[58] K. Higuchi and Y. Horihata, *Weak theories of concatenation and minimal essentially undecidable theories—An encounter of WTC and S2S*. ***Archive for Mathematical Logic***, vol. 53 (2014), no. 7–8, pp. 835–853.

[59] D. Hilbert and P. Bernays, ***Grundlagen der Mathematik, Vols. I and II***, second ed., Springer-Verlag, Berlin, 1934 and 1939.

[60] ———, ***Grundlagen der Mathematik, vol. II***, Springer-Verlag, Berlin, 1939.

[61] D. Isaacson, *Arithmetical truth and hidden higher-order concepts*, ***Logic Colloquium 85*** (J. Barwise, D. Kaplan, H. J. Keisler, P. Suppes, and A. S. Troelstra, editors), Studies in Logic and the Foundations of Mathematics, vol. 122, North-Holland, Amsterdam, 1987, pp. 147–169.

[62] ———, *Necessary and sufficient conditions for undecidability of the Gödel sentence and its truth*, ***Logic, Mathematics, Philosophy: Vintage Enthusiasms*** (D. DeVidi, M. Hallett, and P. Clark, editors), Western Ontario Series in Philosophy of Science, vol. 75, Springer, 2011, pp. 135–152.

[63] T. Jech, *On Gödel's second incompleteness theorem*. ***Proceedings of the American Mathematical Society***, vol. 121 (1994), no. 1, pp. 311–313.

[64] E. Jeřábek, *Recursive functions and existentially closed structures*. ***Journal of Mathematical Logic***, vol. 20 No. 01, 2050002 (2020).

[65] J. P. Jones and J. C. Shepherdson, *Variants of Robinson's essentially undecidable theory. **R***. ***Archive for Mathematical Logic***, vol. 23 (1983), pp. 65–77.

[66] A. Kanamori and K. Mcaloon, *On Gödel's incompleteness and finite combinatorics*. ***Annals of Pure Applied Logic***, vol. 33 (1987), no. 1, pp. 23–41.

[67] R. Kaye and H. Kotlarski, *On models constructed by means of the arithmetized completeness theorem*. ***Mathematical Logic Quarterly***, vol. 46 (2000), no. 4, pp. 505–516.

[68] M. Kikuchi, *A note on Boolos' proof of the incompleteness theorem*. ***Mathematical Logic Quarterly***, vol. 40 (1994), pp. 528–532.

[69] ———, *Kolmogorov complexity and the second incompleteness theorem*. ***Archive for Mathematical Logic***, vol. 36 (1997), no. 6, pp. 437–443.

[70] M. Kikuchi and T. Kurahashi, *Three short stories around Gödel's incompleteness theorems* (*in Japanese*), ***Journal of the Japan Association for Philosophy of Science***, vol. 38 (2011), no. 2, pp. 27–32.

[71] ———, *Generalizations of Gödel's incompleteness theorems for Σ_n-definable theories of arithmetic*. ***The Review of Symbolic Logic***, vol. 10 (2017), no. 4, pp. 603–616.

[72] ———, *Universal Rosser predicates*. ***The Journal of Symbolic Logic***, vol. 82 (2017), no. 1, pp. 292–302.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

[73] M. Kikuchi, T. Kurahashi, and H. Sakai, *On proofs of the incompleteness theorems based on Berry's paradox by Vopěnka, Chaitin, and Boolos*. ***Mathematical Logic Quarterly***, vol. 58 (2012), no. 4–5, pp. 307–316.

[74] M. Kikuchi and K. Tanaka, *On formalization of model-theoretic proofs of Gödel's theorems*. ***Notre Dame Journal of Formal Logic***, vol. 35 (1994), no. 3, pp. 403–412.

[75] L. Kirby, *Flipping properties in arithmetic*. ***The Journal of Symbolic Logic***, vol. 47 (1982), no. 2, pp. 416–422.

[76] S. C. Kleene, *A symmetric form of Godel's theorem*. ***Indagationes Mathematicae***, vol. 12 (1950), pp. 244–246.

[77] H. Kotlarski, *On the incompleteness theorems*. ***The Journal of Symbolic Logic***, vol. 59 (1994), no. 4, pp. 1414–1419.

[78] ———, *Other proofs of old results*. ***Mathematical Logic Quarterly***, vol. 44 (1998), pp. 474–480.

[79] ———, *The incompleteness theorems after* 70 *years*. ***Annals of Pure and Applied Logic***, vol. 126 (2004), no. 1–3, pp. 125–138.

[80] G. Kreisel, *Note on arithmetic models for consistent formulae of the predicate calculus*. ***Fundamenta Mathematicae***, vol. 37 (1950), pp. 265–285.

[81] ———, *On weak completeness of intuitionistic predicate logic*. ***The Journal of Symbolic Logic***, vol. 27 (1962), pp. 139–158.

[82] ———, *A survey of proof theory*. ***The Journal of Symbolic Logic***, vol. 33 (1968), pp. 321–388.

[83] S. Kritchman and R. Raz, *The surprise examination paradox and the second incompleteness theorem*. ***Notices of the American Mathematical Society***, vol. 57 (2010), no. 11, pp. 1454–1458.

[84] G. Leach-Krouse, *Yablifying the Rosser sentence*. ***Journal of Philosophical Logic***, vol. 43 (2014), pp. 827–834.

[85] T. Kurahashi, *Rosser-type undecidable sentences based on Yablo's paradox*. ***Journal of Philosophical Logic***, vol. 43 (2014), pp. 999–1017.

[86] ———, *Arithmetical completeness theorem for modal logic*. ***K***. ***Studia Logica***, vol. 106 (2018), no. 2, pp. 219–235.

[87] ———, *Arithmetical soundness and completeness for* Σ_2 *numerations*. ***Studia Logica***, vol. 106 (2018), no. 6, pp. 1181–1196.

[88] ———, *A note on derivability conditions*. ***The Journal of Symbolic Logic***, 2020, to appear, doi:10.1017/jsl.2020.33.

[89] ———, *Rosser provability and the second incompleteness theorem*. ***Symposium on Advances in Mathematical Logic 2018 Proceedings***, to appear.

[90] M. Li and P. M. B. Vitányi, *Kolmogorov complexity and its applications*, ***Handbook of Theoretical Computer Science*** (J. van Leeuwen, editor), Elsevier, Amsterdam, 1990, pp. 187–254.

[91] P. Lindström, ***Aspects of Incompleteness***, Lecture Notes in Logic, vol. 10, Cambridge University Press, Cambridge, 2017.

[92] M. H. Löb, *Solution of a problem of Leon Henkin*. ***The Journal of Symbolic Logic***, vol. 20 (1955), no. 2, pp. 115–118.

[93] G. Mills, *A tree analysis of unprovable combinatorial statements*. ***Model Theory of Algebra and Arithmetic***, Lecture Notes in Mathematics, vol. 834, Springer, Berlin, 1980, pp. 248–311.

[94] F. Montagna, *On the formulas of Peano arithmetic which are provably closed under modus ponens*. ***Bollettino dell'Unione Matematica Italiana***, vol. 16 (1979), no. B5, pp. 196–211.

[95] A. Mostowski, *A generalization of the incompleteness theorem*. ***Fundamenta Mathematicae***, vol. 49 (1961), pp. 205–232.

[96] ———, *Thirty years of foundational studies*: *Lectures on the development of mathematical logic and the study of the foundations of mathematics in* 1930–1964. ***Acta Philosophica Fennica***, vol. 17 (1965), pp. 1–180.

[97] R. Murawski, ***Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems***, Synthese Library, vol. 286, Springer, Amsterdam, 2013.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

[98] E. Nelson, ***Predicative Arithmetic***, Mathematical Notes, Princeton University Press, Princeton, NJ, 2014.

[99] K. G. Niebergall, *Natural representations and extensions of Gödel's second theorem*, ***Logic Colloquium 01*** (M. Baaz, S. D. Friedman, and J. Krajíček, editors), Lecture Notes in Logic, vol. 20, The Association for Symbolic Logic, 2005, pp. 350–368.

[100] P. Odifreddi, ***Classical Recursion Theory: The Theory of Functions and Sets of Natural Numbers***, Elsevier, Amsterdam, 1992.

[101] L. Pacholski and J. Wierzejewski, ***Model Theory of Algebra and Arithmetic***, Lecture Notes in Mathematics, vol. 834, Springer, Berlin, 1980.

[102] F. Pakhomov, *A weak set theory that proves its own consistency*, preprint, 2019, arXiv:1907.00877v2.

[103] J. Paris and L. Harrington, *A mathematical incompleteness in Peano arithmetic*. ***Handbook of Mathematical Logic*** (J. Barwise, editor), Studies in Logic and Foundations of Mathematics, vol. 90, Elsevier, Amsterdam, 1982, pp. 1133–1142.

[104] J. Paris and L. Kirby, *Accessible independence results for Peano arithmetic*. ***The Bulletin of the London Mathematical Society***, vol. 14 (1982), no. 4, pp. 285–293.

[105] M. B. Pour-El and S. Kripke, *Deduction-preserving "recursive isomorphisms" between theories*. ***Fundamenta Mathematicae***, vol. 61 (1967), pp. 141–163.

[106] G. Priest, *Yablo's paradox*. ***Analysis***, vol. 57 (1997), no. 4, pp. 236–242.

[107] P. Pudlák, *Another combinatorial principle independent of Peano's axioms*, unpublished manuscript, 1979.

[108] ———, *Cuts, consistency statements and interpretations*. ***The Journal of Symbolic Logic***, vol. 50 (1985), pp. 423–441.

[109] ———, *Incompleteness in the finite domain*. ***The Bulletin of Symbolic Logic***, vol. 23 (2017), no. 4, pp. 405–441.

[110] A. Robinson, *On languages which are based on non-standard arithmetic*. ***Nagoya Mathematical Journal***, vol. 22 (1963), pp. 83–117.

[111] J. B. Rosser, *Extensions of some theorems of Gödel and Church*. ***The Journal of Symbolic Logic***, vol. 1 (1936), no. 3, pp. 87–91.

[112] B. Russell, *Mathematical logic as based on the theory of types*. ***American Journal of Mathematics***, vol. 30 (1908), pp. 222–262.

[113] S. Salehi and P. Seraji, *Gödel–Rosser's incompleteness theorem, generalized and optimized for definable theories*. ***Journal of Logic and Computation***, vol. 27 (2017), no. 5, pp. 1391–1397.

[114] ———, *On constructivity and the Rosser property: A closer look at some Gödelean proofs*. ***Annals of Pure and Applied Logic***, vol. 169 (2018), pp. 971–980.

[115] S. G. Simpson, ***Harvey Friedman's Research on the Foundations of Mathematics***, Studies in Logic and the Foundations of Mathematics, vol. 117, Elsevier, Amsterdam, 1985.

[116] ——— (ed.), ***Logic and Combinatorics***, Contemporary Mathematics, vol. 65, American Mathematical Society, Providence, RI, 1987.

[117] ———, ***Subsystems of Second-Order Arithmetic***, Perspectives in Logic, vol. 1, Cambridge University Press, Cambridge, 2009.

[118] P. Smith, ***An Introduction to Gödel's Theorems***, Cambridge University Press, Cambridge, 2013.

[119] C. Smoryński, *The incompleteness theorems*, ***Handbook of Mathematical Logic*** (J. Barwise, editor), Elsevier, Amsterdam, 1982, pp. 821–865.

[120] R. M. Smullyan, ***Gödel's Incompleteness Theorems***, Oxford Logic Guides, vol. 19, Oxford University Press, Oxford, 1992.

[121] ———, ***Diagnolisation and Self-Reference***, Oxford Logic Guides, vol. 27, Clarendon Press, Oxford, 1994.

[122] R. M. Solovay, *Provability interpretations of modal logic*. ***Israel Journal of Mathematics***, vol. 25 (1976), pp. 287–304.

[123] V. Švejdar, *An interpretation of Robinson arithmetic in its Grzegorczyk's weaker variant*. ***Fundamenta Informaticae***, vol. 81 (2007), no. 1–3, pp. 347–354.

[124] A. Tarski and S. Givant, *Tarski's system of geometry*. ***The Bulletin of Symbolic Logic***, vol. 5 (1999), no. 2, pp. 175–214.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

[125] A. Tarski, A. Mostowski, and R. M. Robinson, ***Undecidabe Theories***, Studies in Logic and the Foundations of Mathematics, vol. 13, Elsevier, Amsterdam, 1953.

[126] A. Turing, *Systems of logic based on ordinals*. ***Proceedings of the London Mathematical Society***, vol. 45 (1939), pp. 161–228.

[127] R. L. Vaught, *On a theorem of Cobham concerning undecidable theories*, *Proceedings of the* 1960 *International Congress on Logic, Methodology and Philosophy of Science* (E. Nagel, P. Suppes, and P. Tarski, editors), Stanford University Press, Stanford, 1962, pp. 14–25.

[128] A. Visser, *The provability logics of recursively enumerable theories extending Peano arithmetic at arbitrary theories extending Peano arithmetic*. ***Journal of Philosophical Logic***, vol. 13 (1984), no. 2, pp. 181–212.

[129] ———, *Growing commas*: *A study of sequentiality and concatenation*. ***Notre Dame Journal of Formal Logic***, vol. 50 (2009), no. 1, pp. 61–85.

[130] ———, *Can we make the second incompleteness theorem coordinate free*? ***Journal of Logic and Computation***, vol. 21 (2011), no. 4, pp. 543–560.

[131] ———, *Why the theory **R** is special*. ***Foundational Adventures: Essay in Honour of Harvey Friedman***, College Publications, 2014, pp. 7–23.

[132] ———, *On **Q***. ***Soft Computing***, vol. 21 (2016), pp. 39–56.

[133] ———, *The second incompleteness theorem*: *Reflections and ruminations*, ***Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*** (L. Horsten and P. Welch, editors), Oxford University Press, Oxford, 2016, pp. 67–91.

[134] ———, *The interpretation existence lemma*. ***Feferman on Foundations: Logic, Mathematics, Philosophy***, Outstanding Contributions to Logic, vol. 13, Springer, Cham, Switzerland, 2018, pp. 101–144.

[135] ———, *From Tarski to Gödel*: *Or, how to derive the second incompleteness theorem from the undefinability of truth without self-reference*. ***Journal of Logic and Computation***, vol. 29 (2019), no. 5, pp. 595–604.

[136] ———, *Another look at the second incompleteness theorem*. ***Review of Symbolic Logic***, vol. 13 (2020), no. 2, pp. 269–295.

[137] P. Vopěnka, *A new proof of Gödel's result on non-provability of consistency*. ***Bulletin del'Académie Polonaise des Sciences. Série des Sciences Mathématiques. Astronomiques et Physiques***, vol. 14 (1966), pp. 111–116.

[138] H. Wang, *Undecidable sentences generated by semantic paradoxes*. ***Journal of Symbolic Logic***, vol. 20 (1955), no. 1, pp. 31–43.

[139] A. Weiermann, *An application of graphical enumeration to PA*. ***Journal of Symbolic Logic***, vol. 68 (2003), no. 1, pp. 5–16.

[140] ———, *A classification of rapidly growing Ramsey functions*. ***Proceedings of the American Mathematical Society***, vol. 132 (2004), pp. 553–561.

[141] ———, *Analytic combinatorics, proof-theoretic ordinals, and phase transitions for independence results*. ***Annals of Pure and Applied Logic***, vol. 136 (2005), pp. 189–218.

[142] ———, *Classifying the provably total functions of PA*, this Journal, vol. 12 (2006), no. 2, pp. 177–190.

[143] ———, *Phase transition thresholds for some Friedman-style independence results*. ***Mathematical Logic Quarterly***, vol. 53 (2007), no. 1, pp. 4–18.

[144] D. E. Willard, *Self-verifying axiom systems, the incompleteness theorem and related reflection principles*. ***Journal of Symbolic Logic***, vol. 66 (2001), no. 2, pp. 536–596.

[145] ———, *A generalization of the second incompleteness theorem and some exceptions to it*. ***Annals of Pure and Applied Logic***, vol. 141 (2006), no. 3, pp. 472–496.

[146] S. Yablo, *Paradox without self-reference*. ***Analysis***, vol. 53 (1993), no. 4, pp. 251–252.

[147] R. Zach, *Hilbert's program then and now*. ***Philosophy of Logic***, Handbook of the Philosophy of Science, Elsevier, Amsterdam, 2006, pp. 411–447.

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press

SCHOOL OF PHILOSOPHY
WUHAN UNIVERSITY
WUHAN, 430072 HUBEI, P.R. CHINA
E-mail: world-cyr@hotmail.com

https://doi.org/10.1017/bsl.2020.44 Published online by Cambridge University Press