STUDIA SEMIOTYCZNE (SEMIOTIC STUDIES), t. XXXIV, nr 1 (2020), s. 159–182
ISSN 0137-6608, e-ISSN 2544-073X
DOI: 10.26333/sts.xxxiv1.07
Article

YONG CHENG *

GÖDEL'S INCOMPLETENESS THEOREM AND THE ANTI-MECHANIST ARGUMENT: REVISITED[1]

SUMMARY: This is a paper for a special issue of *Semiotic Studies* devoted to Stanislaw Krajewski's paper (2020). This paper gives some supplementary notes to Krajewski's (2020) on the Anti-Mechanist Arguments based on Gödel's incompleteness theorem. In Section 3, we give some additional explanations to Section 4–6 in Krajewski's (2020) and classify some misunderstandings of Gödel's incompleteness theorem related to Anti-Mechanist Arguments. In Section 4 and 5, we give a more detailed discussion of Gödel's Disjunctive Thesis, Gödel's Undemonstrability of Consistency Thesis and the definability of natural numbers as in Section 7–8 in Krajewski's (2020), describing how recent advances bear on these issues.

KEYWORDS: Gödel's incompleteness theorem, the Anti-Mechanist Argument, Gödel's Disjunctive Thesis, intensionality.

* Wuhan University, School of Philosophy. E-mail: world-cyr@hotmail.com. ORCID: 0000-0003-2408-3886.

[1] 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 in China. I thank the lamp for my feet, and the light on my path. I would also like to thank Peter Koellner for sending me his drafts of (2018a; 2018b) prior to publication, and the referee for detailed helpful comments.

1. Introduction

Gödel's incompleteness theorem is one of the most remarkable and profound discoveries in the 20th century, an important milestone in the history of modern logic. Gödel's incompleteness theorem has wide and profound influence on the development of logic, philosophy, mathematics, computer science and other fields, substantially shaping mathematical logic as well as foundations and philosophy of mathematics from 1931 onward. The impact of Gödel's incompleteness theorem is not confined to the community of mathematicians and logicians, and it has been very popular and widely used outside mathematics.

Gödel's incompleteness theorem raises a number of philosophical questions concerning the nature of mind and machine, the difference between human intelligence and machine intelligence, and the limit of machine intelligence. It is well known that Turing proposed a convincing analysis of the vague and informal notion of "computable" in terms of the precise mathematical notion of "computable by a Turing machine". So we can replace the vague notion of computation with the mathematically precise notion of a Turing machine. In this paper, following Koellner in (2018), we stipulate that the notion "the mind cannot be mechanized" means that the mathematical outputs of the idealized human mind outstrip the mathematical outputs of any Turing machine.[2] A popular interpretation of Gödel's first incompleteness theorem (G1) is that G1 implies that the mind cannot be mechanized. The Mechanistic Thesis claims that the mind can be mechanized. In this paper, we will not examine the broad question of whether the mind can be mechanized, which has been extensively discussed in the literature (e.g. Penrose, 1989; Chalmers, 1995; Lucas, 1996; Lindström, 2006; Feferman, 2009; Shapiro, 1998; 2003; Koellner, 2016; 2018; 2018; Krajewski, 2020). Instead we will only examine the question of whether G1 implies that the mind cannot be mechanized.

This is a paper for a special issue of *Semiotic Studies* devoted to Krajewski's paper (2020). We first give a summary of Krajewski's work in (2020). In (2020), Krajewski gave a detailed analysis of the alleged proof of the nonmechanical, or non-computational, character of the human mind based on Gödel's incompleteness theorem. Following Gödel himself and other leading logicians, Krajewski refuted the Anti-Mechanist Arguments (the Lucas Argument and the Penrose Argument), and claimed that they are not implied by Gödel's incompleteness theorem alone. Moreover, Krajewski (2020) demonstrated the inconsistency of Lucas's arithmetic and the semantic inadequacy of Penrose's arithmetic. Krajewski (2020) also discussed two consequences of Gödel's incompleteness theorem directly related to Anti-Mechanist Arguments: our consistency is not provable (Gödel's Undemonstrability of Consistency Thesis), and we cannot define the

[2] In this paper, we will not consider the performance of actual human minds, with their limitations and defects; but only consider the idealized human mind and look at what it can do in principle (Koellner, 2018a, p. 338).

natural numbers. The discussion in Krajewski's paper is mainly from the philosophical perspective. However, the discussion in this paper is mainly from the logical perspective based on some recent advances on the study of Gödel's incompleteness theorem and Gödel's Disjunctive Thesis. Basically, we agree with Krajewski's analysis of the Anti-Mechanist Arguments and his conclusion that Gödel's incompleteness theorem alone does not imply that the Anti-Mechanist Arguments hold. However, some discussions in (2020) are vague. Moreover, in the recent work on Gödel's Disjunction Thesis one finds precise versions which can actually be proved. The motivation of this paper is to give some supplementary notes to Krajewski's recent paper (2020) on the Anti-Mechanist Arguments based on Gödel's incompleteness theorem.

This paper is structured as follows. In Section 2, we review some notions and facts we will use in this paper. In Section 3, we give some supplementary notes to Section 5–6 in Krajewski's (2020) and classify some misunderstandings of Gödel's incompleteness theorem related to Anti-Mechanist Arguments. In Section 4, we give a more detailed discussion of Gödel's Disjunctive Thesis as in Section 7 in Krajewski's (2020) based on recent advances of the study on Gödel's Disjunctive Thesis in the literature. In Section 5, we give a more precise discussion of Gödel's Undemonstrability of Consistency Thesis and the definability of natural numbers as in Section 8 in Krajewski's paper.

2. Preliminaries

In this section, we review some basic notions and facts used in this paper. Our notations are standard. For textbooks on Gödel's incompleteness theorem, we refer to (Enderton, 2001; Murawski, 1999; Lindström, 1997; Smith, 2007; Boolos, 1993). There are some good survey papers on Gödel's incompleteness theorem in the literature (Smoryński, 1977; Beklemishev, 2010; Kotlarski, 2004; Visser, 2016; Cheng, in press).

In this paper, we focus on first order theory based on countable language, and always assume the arithmetization of the base theory with a recursive set of non-logical constants. For a given theory T, we use $L(T)$ to denote the language of T. For more details about arithmetization, we refer to (Murawski, 1999). Under the arithmetization, any formula or finite sequence of formulas can be coded by a natural number (called the Gödel number of the syntactic item). In this paper, $\ulcorner\varphi\urcorner$ denotes the numeral representing the Gödel number of φ.

We say a set of sentences Σ is *recursive* if the set of Gödel numbers of sentences in Σ is recursive.[3] 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 *essentially undecidable* iff any recursively axiomatizable consistent

[3] For ease of exposition, we will pass back and forth between the two.

extension of T in the same language is undecidable. We say a sentence φ is *independent* of T if $T \nvdash \varphi$ and $T \nvdash \neg\varphi$. 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 $T \vdash \varphi$ or $T \vdash \neg\varphi$). Informally, an interpretation of a theory T in a theory S is a mapping from formulas of T to formulas of S that maps all axioms of T to sentences provable 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. Interpretability can be accepted as a measure of strength of different theories. For the precise definition of interpretation, we refer to (Visser, 2011) for more details.

Theorem 2.1 (Tarski, Mostowski, Robinson, 1953, Theorem 7, p. 22)**.** *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 also essentially undecidable.*

Robinson Arithmetic **Q** was introduced in (1953) by Tarski, Mostowski and Robinson as a base axiomatic theory for investigating incompleteness and undecidability.

Definition 2.2. Robinson Arithmetic **Q** is defined in the language $\{\mathbf{0}, \mathbf{S}, +, \cdot\}$ with the following axioms:

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

The theory **PA** consists of axioms $\mathbf{Q}_1$–$\mathbf{Q}_2$, $\mathbf{Q}_4$–$\mathbf{Q}_7$ in Definition 2.2 and the following axiom scheme of induction:

$$(\varphi(\mathbf{0}) \wedge \forall x(\varphi(x) \rightarrow \varphi(\mathbf{S}x))) \rightarrow \forall x \varphi(x),$$

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 **PA**. We say $\varphi \in L(\mathbf{PA})$ is a true sentence of arithmetic if $\mathfrak{N} \models \varphi$. We define that $Th(\mathbb{N}, +, \cdot)$ is the set of sentence φ in $L(\mathbf{PA})$ such that $\mathfrak{N} \models \varphi$. Similarly, we have the definition of $Th(\mathbb{Z}, +, \cdot)$, $Th(\mathbb{Q}, +, \cdot)$ and $Th(\mathbb{R}, +, \cdot)$.

We introduce a hierarchy of $L(\mathbf{PA})$-formulas called the "arithmetical hierarchy" (Murawski, 1999; Hájek, Pudlák, 1993). Bounded formulas (Σ^0_0, or Π^0_0, or

Δ_0^0 formula) 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_{n+1}^0 if it has the form $\exists x\varphi$ where φ is Π_n^0. A formula is Π_{n+1}^0 if it has the form $\forall x\varphi$ where φ is Σ_n^0. Thus, a Σ_n^0-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_n^0-formulas. A formula is Δ_n^0 if it is equivalent to both a Σ_n^0 formula and a Π_n^0 formula.

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 $n \in \omega$, $T \vdash \neg\varphi(\bar{n})$. A theory T is 1-consistent if there is no such formula $\varphi(x)$ which is Δ_1^0. A theory T is sound iff for any formula φ, if $T \vdash \varphi$, then $\mathfrak{N} \models \varphi$; a theory T is Σ_1^0-sound iff for any Σ_1^0 formula φ, if $T \vdash \varphi$, then $\mathfrak{N} \models \varphi$.

In the following, unless stated otherwise, let T be a recursively axiomatizable consistent extension of **PA**. There is a formal arithmetical formula $\mathbf{Proof}_T(x,y)$ (called Gödel's proof predicate) which represents the recursive relation $Proof_T(x,y)$ saying that y is the Gödel number of a proof in T of the formula with Gödel number x. Define $\mathbf{Prov}_T(x) \triangleq \exists y\mathbf{Proof}_T(x,y)$. Since we will discuss general provability predicates based on proof predicates, now we give a general definition of proof predicate which is a generalization of properties of Gödel's proof predicate $\mathbf{Proof}_T(x,y)$.

Definition 2.3. We say a formula $\mathbf{Prf}_T(x,y)$ is a proof predicate of T if it satisfies the following conditions:[4]

(1) $\mathbf{Prf}_T(x,y)$ is $\Delta_1^0(\mathbf{PA})$;[5]

(2) $\mathbf{PA} \vdash \forall x(\mathbf{Prov}_T(x) \leftrightarrow \exists y\mathbf{Prf}_T(x,y))$;

(3) for any $n \in \omega$ and formula φ, $\mathbb{N} \models \mathbf{Proof}_T(\ulcorner\varphi\urcorner, n) \leftrightarrow \mathbf{Prf}_T(\ulcorner\varphi\urcorner, n)$;

(4) $\mathbf{PA} \vdash \forall x\forall x'\forall y(\mathbf{Prf}_T(x,y)\wedge\mathbf{Prf}_T(x',y) \rightarrow x = x')$.

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 $\neg\mathbf{Pr}_T(\ulcorner\mathbf{0} \neq \mathbf{0}\urcorner)$.

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

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

D3: $T \vdash \mathbf{Pr}_T(\ulcorner\varphi\urcorner) \rightarrow \mathbf{Pr}_T(\ulcorner\mathbf{Pr}_T(\ulcorner\varphi\urcorner)\urcorner)$.

---

[4] 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".

[5] We say a formula φ is $\Delta_1^0(\mathbf{PA})$ if there exists a Σ_1^0 formula α such that $\mathbf{PA} \vdash \varphi \leftrightarrow \alpha$, and there exists a Π_1^0 formula β such that $\mathbf{PA} \vdash \varphi \leftrightarrow \beta$.

D1–**D3** is called the Hilbert-Bernays-Löb derivability condition. Note that **D1** holds for any provability predicate $\mathbf{Pr}_T(x)$. We say that provability predicate $\mathbf{Pr}_T(x)$ is standard if it satisfies **D2** and **D3**. In this paper, unless stated otherwise, we assume that $\mathbf{Con}(T)$ is the canonical arithmetic sentence expressing the consistency of T and $\mathbf{Con}(T)$ is formulated via a standard provability predicate.

The reflection principle for T, denoted by $\mathbf{Rfn}_T$, is the schema $\mathbf{Pr}_T(\ulcorner\varphi\urcorner) \rightarrow \varphi$ for every sentence φ in $L(T)$. The reflection principle for T restricted to a class of sentences Γ will be denoted by Γ-$\mathbf{Rfn}_T$.

Let $\alpha(x)$ be a formula in $L(T)$. We can similarly define the provability predicate and consistency statement w.r.t. formula $\alpha(x)$ as follows. 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 $\neg\mathbf{Pr}_\alpha(\ulcorner\mathbf{0} \neq \mathbf{0}\urcorner)$. We say that formula $\alpha(x)$ is a numeration of T if for any n, $T \vdash \alpha(\overline{n})$ iff n is the Gödel number of some sentence in T.

3. Some notes on Gödel-Based Anti-Mechanist Arguments

There has been a massive amount of literature on the Anti-Mechanist Arguments due primarily to Lucas and Penrose (see Lucas, 1961; Penrose; 1989) which claim that G1 shows that the human mind cannot be mechanized. The Anti-Mechanist Argument began with Nagel and Newman in (2001) and continued with Lucas's publication in (1961). Nagel and Newman's argument was criticized by Putnam in (1960) and earlier by Gödel (Feferman, 2009), while Lucas's argument was much more widely criticized in the literature. See Feferman (2009) for a historical account and Benacerraf (1967) for an influential criticism of Lucas. Penrose proposed a new argument for the Anti-Mechanist Argument in (1994; 2011). Penrose's new argument is the most sophisticated and promising Anti-Mechanist Argument which has been extensively discussed and carefully analyzed in the literature (Chalmers, 1995; Feferman, 1995; Lindström, 2001; 2006; Shapiro, 1998; 2003; Gaifman, 2000; Koellner, 2016; 2018a; 2018b, etc.)

Most philosophers and logicians believe that variants of the arguments of Lucas and Penrose are not fully convincing. However, they do not agree so well on what is wrong with arguments of Lucas and Penrose. One strength of Krajewski's paper (2020) is that it provides a detailed review of the history of Anti-Mechanist Arguments based on Gödel's incompleteness theorem (Krajewski, 2020, Section 3) and an analysis of these Gödel-Based Anti-Mechanist Arguments (e.g. Lucas's argument in Section 4 and Penrose's argument in Section 6 in [Krajewski, 2020]). In this section, based on Krajewski's work, we give some supplementary notes of Krajewski's Sections 5–6.

For us, the Gödel-Based Anti-Mechanist Argument comes from some misinterpretations of Gödel's incompleteness theorem. To understand the source of these misinterpretations or illusions, we should first have correct interpretations of Gödel's incompleteness theorem. In the following, we first review some important facts about Gödel's incompleteness theorem which are helpful to clarify some misinterpretations of Gödel's incompleteness theorem.

Gödel proved his incompleteness theorem in (1931) for a certain formal system **P** related to Russell-Whitehead's Principia Mathematica and based on the simple theory of types over the natural number series and the Dedekind-Peano axioms (Beklemishev, 2010, p. 3). Gödel's original first incompleteness theorem (1931, Theorem VI) says that for formal theory T formulated in the language of **P** and obtained by adding a primitive recursive set of axioms to the system **P**, if T is ω-consistent, then T is incomplete. The following theorem is a modern reformulation of Gödel's first incompleteness theorem.

Theorem 3.1 (Gödel's first incompleteness theorem (G1))**.** *If T is a recursively axiomatized extension of* **PA**, *then there exists a Gödel sentence* **G** *such that:*

(1) if T is consistent, then $T \nvdash$ **G***;*

*(2) if T is ω-consistent, then $T \nvdash \neg$***G**.

Thus if T is ω-consistent, then **G** is independent of T and hence T is incomplete. If T is consistent, Gödel sentence **G** is a true Π^0_1 sentence of arithmetic. Gödel's proof of G1 is constructive: one can effectively find a true Π^0_1 sentence **G** of arithmetic such that **G** is independent of T assuming T is ω-consistent. Gödel calls this the "incompletability or inexhaustability of mathematics". Note that only assuming that T is consistent, we can show that **G** is a true sentence of arithmetic unprovable in T. But it is not enough to show that $T \nvdash \neg$**G** only assuming that T is consistent. To show that $T \nvdash \neg$**G**, we need a stronger condition such as "T is 1-consistent" or "T is Σ^0_1-sound".

Let T be a recursively axiomatized extension of **PA**. After Gödel, Rosser constructed Rosser sentence **R** (a Π^0_1 sentence) and showed that if T is consistent, then **R** is independent of T. Rosser improved Gödel's G1 in the sense that Rosser proved that T is incomplete only assuming that "T is consistent" which is weaker than "T is 1-consistent".

In this paper, let $\langle M_n : n \in \omega\rangle$ be the list of Turing machines and $Th(M_n)$ be the set of sentences produced by the Turing machine M_n. Let $C = \{n : Th(M_n)$ is a consistent theory$\}$ and $S = \{n : Th(M_n)$ is a sound theory$\}$.

The following proposition on inconsistency and unsoundness is from (Krajewski, 2020).

Proposition 3.2.

(1) If F is a partial recursive function such that $C \subseteq dom(F)$ and $F(n) \notin Th(M_n)$ for any $n \in C$, then $\{F(n) : n \in dom(F)\}$ is inconsistent.

(2) If F is a partial recursive function such that $S \subseteq dom(F)$ and $F(n) \notin Th(M_n)$ for any $n \in S$, then $\{F(n) : n \in dom(F)\}$ is inconsistent.

A natural question is: whether there exists such a function F with these properties. However, the effective version of Gödel's first incompleteness theorem (EG1) tells us that there exists a partial recursive function F such that for any $n \in \omega$, if $Th(M_n)$ is consistent, then $F(n)$ is defined and $F(n)$ is the Gödel number of a true arithmetic sentence which is not provable in $Th(M_n)$. Thus there exists such a function F with the properties as stated in Proposition 3.2.

One popular interpretation of EG1 is: for any Turing machine M_n, $F(n)$ picks up the true sentence of arithmetic not produced by M_n. However, this is a misinterpretation of EG1 which in fact says that for such a partial recursive function F, if $Th(M_n)$ is consistent, then $F(n)$ is the Gödel number of a true sentence of arithmetic which is not provable in $Th(M_n)$. A natural question is: whether there exists an effective procedure such that we can decide whether $Th(M_n)$ is consistent. The answer is negative since C is a complete Π^0_1 set as Koellner points out in (2018a).

Krajewski (2020) claimed that C and S are not recursive. However, as Krajewski (2020) commented, Proposition 3.2 on inconsistency and unsoundness does not require that for $n \in dom(F)$, $F(n)$ is the code of a true arithmetic sentence. But we do not see that C or S is not recursive from Proposition 3.2. However, if we add the condition that for $n \in dom(F)$, $F(n) \in \mathbf{Truth} \setminus Th(M_n)$, then we can show that C and S are not recursive. Let us take C for example and show that C is not recursive.

Proposition 3.3. *C is not recursive.*[6]

Proof. Suppose C is recursive. Let $A = \{F(n) : n \in C\}$. Then A is recursive enumerable. Suppose $A = Th(M_m)$ for some m. Note that $A \subseteq \mathbf{Truth}$, and so A is consistent. By the definition of C, $m \in C$ and hence $F(m) \in A$. But, on the other hand, $F(m) \notin Th(M_m) = A$ which leads to a contradiction. □

Since C is undecidable, it is impossible to effectively distinguish the case that $Th(M_n)$ is consistent and the case that $Th(M_n)$ is not consistent.

In fact, Theorem 3.2 can be generalized in the following form:

Theorem 3.4. *Let P be any property about first order theory (i.e. consistency, soundness, 1-consistency, etc). Let $C = \{n : Th(M_n)$ has property $P\}$. Suppose F is a partial recursive function satisfying the following conditions:*

(1) $C \subseteq dom(F)$,

[6] In fact, C is a complete Π^0_1 set as Koellner points out in (2018a).

(2) for each $n \in C$, $F(n) \notin Th(M_n)$.

Then, $\{F(n) : n \in dom(F)\}$ *does not have property P.*

Proof. Let $A = \{F(n) : n \in dom(F)\}$. Suppose A has property P. Since F is partial recursive, A is recursively enumerable. Suppose $A = Th(M_k)$ for some k. Since A has property P, we have $k \in C$. Thus, $F(k) \notin Th(M_k) = A$ which contradicts that $F(k) \in A$. □

Gödel announced 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 Zach, 2007, p. 431). For a theory T, recall that $\mathbf{Con}(T)$ is the canonical arithmetic sentence expressing the consistency of T under Gödel's recursive arithmetization of T. The following is a modern reformulation of G2:

Theorem 3.5. *Let T be a recursively axiomatized extension of* **PA**. *If T is consistent, then* $T \nvdash \mathbf{Con}(T)$.

From G2, we cannot get that $\mathbf{Con}(T)$ is independent of T only assuming that T is consistent. It is provable in T that if T is consistent, then $T \vdash \mathbf{Con}(T) \leftrightarrow \mathbf{G}$ and thus $T \nvdash \mathbf{Con}(T)$. However, it is not provable in T that if T is consistent, then $T + \mathbf{Con}(T)$ is also consistent.[7] So it is not enough to show that $T \nvdash \neg\mathbf{Con}(T)$ only assuming that T is consistent. But we could prove that $\mathbf{Con}(T)$ is independent of T by assuming that T is 1-consistent which is stronger than the condition "T is consistent".[8] Let 1-$\mathbf{Con}(T)$ be the sentence in $L(\mathbf{PA})$ expressing that T is 1-consistent. Fact 3.6 is a summary of these results.

Fact 3.6. Let T be a recursively axiomatized consistent extension of **PA**.

(1) $T \vdash \mathbf{Con}(T) \rightarrow \mathbf{Con}(T + \neg\mathbf{Con}(T))$;

(2) $T \nvdash \mathbf{Con}(T) \rightarrow \mathbf{Con}(T + \mathbf{Con}(T))$;

(3) $T \vdash \mathbf{Con}(T) \rightarrow \mathbf{Con}(T + \mathbf{R})$;[9]

(4) $T \vdash$ 1-$\mathbf{Con}(T) \rightarrow \mathbf{Con}(T + \mathbf{Con}(T))$.

An illusion of the application of Gödel's incompleteness theorem is that we can add consistencies (or Out-Gödeling) forever: from $\mathbf{Con}(T)$, we have

[7] See (Boolos, 1993, Theorem 4, p. 97) for a modal proof in **GL** of this fact using the arithmetic completeness theorem for **GL**.

[8] 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.

[9] Recall that **R** is the Rosser sentence.

$\mathbf{Con}(T + \mathbf{Con}(T))$, then $\mathbf{Con}(T + \mathbf{Con}(T + \mathbf{Con}(T)))$ and so on. However, by Fact 3.6, this does not hold. For the iteration of adding the consistency statement (or Out-Gödeling), we need a stronger condition: T is 1-consistent. The following fact shows the difference between $\mathbf{Con}(T)$ and 1-$\mathbf{Con}(T)$.

Fact 3.7 (Smoryński, 1977)**.** Let T be a recursively axiomatized consistent extension of **PA**. Then $T \vdash \mathbf{Con}(T) \leftrightarrow \Pi^0_1\text{-}\mathbf{Rfn}_T$ and $T \vdash 1\text{-}\mathbf{Con}(T) \leftrightarrow \Sigma^0_1\text{-}\mathbf{Rfn}_T$.

As a corollary of Fact 3.7, $1\text{-}\mathbf{Con}(T) \vdash 1\text{-}\mathbf{Con}(T + \mathbf{Con}(T))$ (see Proposition 3 in Pudlák, 1999). Thus, if we assume 1-$\mathbf{Con}(T)$, then we can prove $\mathbf{Con}(T)$, $\mathbf{Con}(T + \mathbf{Con}(T))$, $\mathbf{Con}(T + \mathbf{Con}(T + \mathbf{Con}(T)))$ and we can continue forever (note that the assumption 1-$\mathbf{Con}(T)$ is stronger than all these statements).

In summary, the differences between Rosser sentence and Gödel sentence, as well as between $\mathbf{Con}(T)$ and 1-$\mathbf{Con}(T)$ are very important. However, these differences are often overlooked in informal philosophical discussions of Gödel's incompleteness theorem.

4. Gödel's Disjunctive Thesis

The focus of Krajewski's paper (2020) is not about Gödel's Disjunctive Thesis even if he gives a very brief discussion of Gödel's Disjunctive Thesis related to the Anti-Mechanist Arguments in Section 7. In this section, we give a more detailed discussion of Gödel's Disjunctive Thesis and its relevance to the Mechanistic Thesis based on recent advances on the study of Gödel's Disjunctive Thesis. This section is a summary of Koellner's papers (2018a) and (2018b), and we follow Koellner's presentation very closely.

Gödel did not argue that his incompleteness theorem implies that the mind cannot be mechanized. Instead, Gödel argued that his incompleteness theorem implies a weaker conclusion: Gödel's Disjunctive Thesis (GD).

The first disjunct: The mind cannot be mechanized.

The second disjunct: There are absolutely undecidable statements.[10]

Gödel's Disjunctive Thesis (GD)**:** Either the first disjunct or the second disjunct holds.[11]

[10] In the sense that there are mathematical truths that cannot be proved by the idealized human mind.

[11] The original version of GD was introduced by Gödel in (1951; see p. 310): "So the following disjunctive conclusion is inevitable: either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say, the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified (where the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives)".

Gödel's Disjunctive thesis (GD) concerns the limit of mathematical knowledge and the possibility of the existence of mathematical truths that are inaccessible to the idealized human mind. The first disjunct expresses an aspect of the power of the idealized human mind, while the second disjunct expresses an aspect of its limitations.[12]

What about Gödel's view toward the first disjunct and the second disjunct? For Gödel, the first disjunct is true and the second disjunct is false; that is the mind cannot be mechanized and human mind is sufficiently powerful to capture all mathematical truths. Gödel's incompleteness theorem shows certain weaknesses and limitations of one given Turing machine. For Gödel, mathematical proof is an essentially creative activity and his incompleteness theorem indicates the creative power of human reason. Gödel believes that the distinctiveness of the human mind when compared to a Turing machine is evident in its ability to come up with new axioms and develop new mathematical theories. Gödel shared Hilbert's belief expressed in 1926 in the words: "in mathematics there is no ignoramuses, we should know and we must know" (see Reid, 1996, p. 192). Based on his rationalistic optimism, Gödel believed that we are arithmetically omniscient and the second disjunct is false.[13] However, Gödel admits that he cannot give a convincing argument for either the first disjunct or the second disjunct. Gödel thinks that the most he can claim to have established is his Disjunctive Thesis. For Gödel, GD is a "mathematically established fact" of great philosophical interest which follows from his incompleteness theorem, and it is "entirely independent from the standpoint taken toward the foundation of mathematics" (Gödel, 1951, p. 310).[14] In the following, we give a concise overview of the current progress on Gödel's disjunctive thesis based on Koellner's work in (2016; 2018a; 2018b).

Let **K** be the set of sentences in $L(\mathbf{PA})$ that the idealized human mind can know. Let **Truth** be the set of sentences in $L(\mathbf{PA})$ which are true in the standard model of arithmetic and **Prov** be the set of sentences in $L(\mathbf{PA})$ which are provable in **PA**. Gödel refers to **Truth** as objective mathematics and **K** as subjective mathematics. Recall that a theory T in $L(\mathbf{PA})$ is sound if $T \subseteq \mathbf{Truth}$. In this paper, we assume that **K** is sound. However, from G1, we have $\mathbf{Prov} \subsetneq \mathbf{Truth}$ since Gödel's sentence is a true sentence of arithmetic not provable in **PA**.[15]

[12] We refer to (Horsten, Welch, 2016), a recent comprehensive research volume about GD, for more discussions of the status of GD.

[13] For more discussions of the status of the second disjunct, we refer to (Horsten, Welch, 2016).

[14] In the literature there is a consensus that Gödel's argument for GD is definitive, but until now we have no compelling evidence for or against any of the two disjuncts (Horsten, Welch, 2016).

[15] Let us take Fermat's last theorem for another example. People have shown that Fermat's last theorem is a true sentence of arithmetic but, as far as I know, it is still an open problem whether Fermat's last theorem is provable in **PA**. So Fermat's last theorem belongs to **K** but it is open whether it belongs to **Prov**.

Note that GD concerns the concepts of relative provability, absolute provability, and truth. Before we present the analysis of GD, let us first examine two key notions about provability: relative provability and absolute provability. The notion of relative provability is well understood and we have a precise definition of relative provability in a formal system. But the notion of absolute provability is much more ambiguous and we have no unambiguous formal definition of absolute provability as far as we know. The notion of absolute provability is intended to be intensionally different from the notion of relative provability in that absolute provability is not conceptually connected to a formal system. In contrast to the notion of relative provability, there is little agreement on what principles of the notion "absolute provability" should be adopted. In this paper, we identify the notion of "relatively provable with respect to a given formal system F" with the notion of "producible by a Turing machine M" (where M is the Turing machine corresponding to F)[16] and we identify the notion of "absolute provability" with the notion of "what the idealized human mind can know".[17] Under this assumption, **K** is just the set of sentences that are absolutely provable.

In this paper, we assume without loss of generality that $\mathbf{Q} \subseteq Th(M_n)$ such that both G1 and G2 apply to $Th(M_n)$. For a natural number n, we say that a statement φ is *relatively undecidable* w.r.t. theory $Th(M_n)$ for some n if $\varphi \notin Th(M_n)$ and $\neg\varphi \notin Th(M_n)$. We say that a statement φ is *absolutely undecidable* if $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$. Let us first examine what the incompleteness theorem tells us about the relationship between $Th(M_n)$, **K** and **Truth**.

Note that G1 tells us that for any sufficiently strong consistent theory F containing **Q**, there are statements which are relatively undecidable with respect to F. But as Gödel argued, these statements are not absolutely undecidable; instead one can always pass to higher systems in which the sentence in question is provable (see Gödel, 1995, p. 35). For example, from G2, **Con**(**PA**) is not provable in **PA**; but **Con**(**PA**) is provable in second order arithmetic ($\mathbf{Z_2}$). Since G2 applies to $\mathbf{Z_2}$, the Π^1_0-truth **Con**($\mathbf{Z_2}$) is not provable in $\mathbf{Z_2}$. But **Con**($\mathbf{Z_2}$) is provable in $\mathbf{Z_3}$ (third order arithmetic) which captures the Π^1_0-truth that was missed by $\mathbf{Z_2}$. This pattern continues up through the orders of arithmetic and up through the hierarchy of set-theoretic systems; at each stage a missing Π^1_0-truth is captured at the next stage (see Koellner, 2018a, p. 347).

Now let us examine the question of whether the incompleteness theorem shows that GD holds. From the literature, we have found a natural framework $\mathbf{EA_T}$ in which we can show that if the concepts of relative provability, absolute provability and truth satisfy some principles, then one can give a rigorous proof of GD, vindicating Gödel's claim that GD is a mathematically established fact (see Koellner, 2018a, p. 355).

[16] Note that sentences relatively provable with respect to a given formal system F can be enumerated by a Turing machine.

[17] Williamson (2016) makes the similar definition that a mathematical hypothesis is absolutely decidable if and only if either it or its negation can in principle be known by a normal mathematical process; otherwise it is absolutely undecidable.

Now we introduce two systems of epistemic arithmetic: **EA** and **EA$_\mathbf{T}$**. For the presentation of **EA** and **EA$_\mathbf{T}$**, we closely follow Koellner's discussion in (2016; 2018a). The first is designed to deal with $Th(M_e)$ and **K**, and the second is designed to deal with $Th(M_e)$, **K** and **Truth**. For **EA$_\mathbf{T}$**, we only require a typed truth predicate.[18] The basic system **EA** of epistemic arithmetic has axioms of arithmetic and axioms of absolute provability, and the extended system **EA$_\mathbf{T}$** has additional axioms of typed truth.[19] In **EA** and **EA$_\mathbf{T}$**, **K** is treated as an operator rather than a predicate. From results in Gödel (1986), Myhill (1960), Montague (1963), Thomason (1980), and others, if one formulates a theory of absolute provability with **K** as a predicate then inconsistency may come (see Koellner, 2016). The basic axioms of absolute provability are:[20]

K1: Universal closures of formulas of the form $\mathbf{K}\varphi$ where φ is a first-order validity.

K2: Universal closures of formulas of the form $(\mathbf{K}(\varphi \to \psi) \wedge \mathbf{K}\varphi) \to \mathbf{K}\psi$.

K3: Universal closures of formulas of the form $\mathbf{K}\varphi \to \varphi$.

K4: Universal closures of formulas of the form $\mathbf{K}\varphi \to \mathbf{KK}\varphi$.[21]

The language L(**EA**) is L(**PA**) expanded to include an operator **K** that takes formulas of L(**EA**) as arguments. The axioms of arithmetic are simply those of **PA**, only now the induction scheme is taken to cover all formulas in L(**EA**). For a collection Γ of formulas in L(**EA**), let $\mathbf{K}\Gamma$ denote the collection of formulas $\mathbf{K}\varphi$ where $\varphi \in \Gamma$. The system **EA** is the theory axiomatized by $\Sigma \cup \mathbf{K}\Sigma$, where Σ consists of the axioms of **PA** in the language L(**EA**) and the basic axioms of absolute provability. The language L(**EA$_\mathbf{T}$**) of **EA$_\mathbf{T}$** is the language L(**EA**) augmented with a unary predicate T. The system **EA$_\mathbf{T}$** is the theory axiomatized by $\Sigma \cup \mathbf{K}\Sigma$, where Σ consists of the axioms of **PA** in the language L(**EA$_\mathbf{T}$**), the basic

[18] A typed truth predicate is one that applies only to statements that do not themselves involve the truth predicate. In contrast, a type-free truth predicate is one which also applies to statements that themselves involve the truth predicate. The principles governing typed truth predicates are perfectly straightforward and uncontroversial, while the principles governing type-free truth predicates are much more delicate (Koellner, 2018a).

[19] These systems were first introduced by Myhill (1960), Reinhardt (1985a; 1985b; 1986) and Shapiro (1985), and then investigated by many others (e.g. Horsten, 1998; Leitgeb, 2009; Carlson, 2000; Koellner, 2016; 2018a and others).

[20] The basic conditions we will impose on knowability are: (1) if the idealized human mind knows φ and $\varphi \to \psi$ then the idealized human mind knows ψ; (2) if the idealized human mind knows φ then φ is true; (3) if the idealized human mind knows φ then the idealized human mind knows that the idealized human mind knows φ.

[21] **K1**-known as logical omniscience-says that **K** holds of all first-order logical validities; **K2** says that **K** is closed under modus ponens, and so distributes across logical derivations; **K3** says that **K** is correct; and **K4** says that **K** is absolutely self-reflective (Koellner, 2018a).

axioms of absolute provability (in the language $L(\mathbf{EA_T})$), and the Tarskian axioms of truth for the language $L(\mathbf{EA})$.

From the incompleteness theorem, Gödel made the following two claims about the relationship between $Th(M_e)$, $\mathbf{K}$ and $\mathbf{Truth}$.

Claim One: For any $e \in \mathbb{N}$, $\mathbf{K}(Th(M_e) \subseteq \mathbf{Truth}) \rightarrow Th(M_e) \subsetneq \mathbf{K}$.[22]

Claim Two: Either $\neg\exists e(Th(M_e) = \mathbf{K})$ or $\exists\varphi(\varphi \in \mathbf{Truth} \wedge \varphi \notin \mathbf{K} \wedge \neg\varphi \notin \mathbf{K})$.[23]

Gödel's Claim One is formalizable and provable in $\mathbf{EA_T}$. In fact, something stronger is provable in **EA** as the following theorem shows:

Theorem 4.1 (Reinhardt, 1985a)**.** *Assume that S includes* **EA**. *Suppose F*(*x*) *is a formula with one free variable.*

(1) If for each sentence φ, $S \vdash \mathbf{K}(F(\ulcorner\varphi\urcorner) \rightarrow \varphi)$. *Then there is a sentence* ϕ *such that* $S \vdash \mathbf{K}\phi \wedge \mathbf{K}\neg F(\ulcorner\phi\urcorner)$.

(2) If for each sentence φ, $S \vdash \mathbf{K}(\mathbf{K}\varphi \rightarrow F(\ulcorner\varphi\urcorner))$. *Then* $S \vdash \mathbf{K}\neg\mathbf{K}(\mathbf{Con}(F))$.

From the following theorem, GD is also formalizable and provable in $\mathbf{EA_T}$ which confirms Gödel's claim that GD is a mathematically established fact.[24]

Theorem 4.2 (Reinhardt, 1986)**.** *Assume* $\mathbf{EA_T}$. *Then* GD *holds.*

Following Reinhardt, we should distinguish three levels of the mechanistic thesis.

(1) The weak mechanistic thesis (WMT): $\exists e(\mathbf{K} = Th(M_e))$;

(2) The strong mechanistic thesis (SMT): $\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$;

(3) The super strong mechanistic thesis (SSMT): $\exists e\,\mathbf{K}(\mathbf{K} = Th(M_e))$.

Note that WMT is just the first disjunct which says that there is a Turing machine which coincides with the idealized human mind in the sense that the two have the same outputs. Note that SMT says that the idealized human mind knows that

[22] The informal proof of Claim One is as follows: Suppose $\mathbf{K}(Th(M_e) \subseteq \mathbf{Truth})$. Since it is knowable that $Th(M_e)$ is consistent, it is knowable that there is a true sentence of arithmetic which is not provable in $Th(M_e)$. So $Th(M_e) \subsetneq \mathbf{K}$.

[23] The informal proof of Claim Two is as follows: Suppose $Th(M_e) = \mathbf{K}$ for some e. Since $Th(M_e)$ is R.E. but Truth is not arithmetic, $\mathbf{K} \subsetneq \mathbf{Truth}$. So we can find some $\varphi \in \mathbf{Truth}$ but $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$.

[24] It is a little delicate to formalize GD in $\mathbf{EA_T}$ since $\mathbf{K}$ is formalized as an operator in $\mathbf{EA_T}$ and so we are prohibited from quantifying into it. For the details, we refer to Reinhardt (1986) and Koellner (2016; 2018a).

there is a Turing machine which coincides with the idealized human mind. Note that SSMT says that there is a particular Turing machine such that the idealized human mind knows that that particular machine coincides with the idealized human mind.

Suppose WMT holds. Then there exists an e^* such that in fact $\mathbf{K} = Th(M_{e*})$. It might seem at first that if we know that there is such an e^* then we will be able to find, in a computable way, the indices e such that $\mathbf{K} = Th(M_e)$. But this is an illusion, as demonstrated by Rice's Theorem, which we shall now explain.

In recursion theory, the sets $Th(M_e)$ are known as *computably enumerable* sets. Each such set is the domain of a partial computable function φ_e. Rice's Theorem states that for any class C of partial computable functions, $\{e : \varphi_e \in C\}$ is computable iff either $C = \emptyset$ or C is the class of all partial computable functions. Now consider the set of indices that we are interested in, namely, $\{e : \mathbf{K} = dom(\varphi_e)\}$, that is, $\{e : \varphi_e \in C\}$ where $C = \{\varphi_e : \mathbf{K} = dom(\varphi_e)\}$. It follows immediately from Rice's theorem that $\{e : \mathbf{K} = dom(\varphi_e)\}$ is not computable.

The following theorem shows that we can prove in $\mathbf{EA_T}$ that there does not exist a particular Turing machine such that the idealized human mind knows that that particular Turing machine coincides with the idealized human mind.

Theorem 4.3 (Reinhardt, 1985a)**.** $\mathbf{EA_T}$ + SSMT *is inconsistent.*

The following theorem shows that, from the viewpoint of $\mathbf{EA_T}$ it is possible that the idealized human mind is in fact a Turing machine. From Theorem 4.3, it just cannot know which one.

Theorem 4.4 (Reinhardt, 1985b)**.** $\mathbf{EA_T}$ + WMT *is consistent.*

From Theorem 4.4, the first disjunct is not provable in $\mathbf{EA_T}$. But Gödel did think that one day we would be in a position to prove the first disjunct, and what was missing, as he saw it, was an adequate resolution of the paradoxes involving self-applicable concepts like the concept of truth. Gödel thought that "[i]f one could clear up the intensional paradoxes somehow, one would get a clear proof that mind is not machine".[25]

The following technical theorem from Carlson shows that, from the point of view of $\mathbf{EA_T}$, it is possible that the idealized human mind knows that it is a Turing machine: it just cannot know which one.

Theorem 4.5 (Carlson, 2000)**.** $\mathbf{EA_T}$ + SMT *is consistent.*

Now we give a summary for the question whether Gödel's incompleteness theorems imply the first disjunct. The incompleteness theorems imply that

[25] This quotation is from Hao Wang's reconstruction of his conversations with Gödel (see Wang, 1996, p. 187).

$\neg\exists e\,\mathbf{K}(\mathbf{K} = Th(M_e))$. But from Theorem 4.4, it does not follow that $\neg\exists e(\mathbf{K} = Th(M_e))$; and from Theorem 4.5, it does not even follow that $\neg\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$. The difference between $\exists e\,\mathbf{K}$ and $\mathbf{K}\exists e$ before $\mathbf{K} = Th(M_e)$ is essential. Assuming the principles embodied in $\mathbf{EA_T}$, it is possible to know that we are a Turing machine (i.e. $\mathbf{K}\exists e(\mathbf{K} = Th(M_e))$); it is just not possible for there to be a Turing machine such that we know that we are that Turing machine (i.e. $\exists e\,\mathbf{K}(\mathbf{K} = Th(M_e))$).

Penrose proposed a new argument for the first disjunct in (1994, 2011). Penrose's new argument is the most sophisticated and promising argument for the first disjunct. It has been extensively discussed and carefully analyzed in the literature (see Chalmers, 1995; Feferman, 1995; Lindström, 2001; 2006; Shapiro, 1998; 2003; Gaifman, 2000; Koellner, 2016; 2018b, etc). The question of whether Penrose's new argument establishes the first disjunct is quite subtle. Penrose's new argument involves treating truth as type-free, and so for the analysis and formalization of Penrose's new argument, we need to employ type-free notions of truth. However, we now have many type-free theories of truth and there is no consensus as to which option is best. Koellner was the first to discuss Penrose's new argument in the context of type-free truth. And he shows that when one shifts to a type-free notion of truth then one can treat **K** as a predicate (as a contrast, in the context of **EA** and $\mathbf{EA_T}$, **K** cannot be treated as a predicate).

In the literature, Koellner proposed the framework **DTK** which employs Feferman's type-free theory of determinate truth **DT** and some additional axioms governing **K** to the axioms of **DT**.[26] The following results about the system **DTK** are due to Koellner. From (Koellner, 2016; 2018b), **DTK** is consistent (see 2016, Theorem 7.14.1) and **DTK** proves GD (see 2016, Theorem 7.15.3). However, the particular argument Penrose gives for the first disjunct fails in the context of **DTK** (see 2018b, Theorem 4.1). Moreover, even if we restrict the first and second disjunct to arithmetic statements, **DTK** can neither prove nor refute either the first disjunct or the second disjunct (see 2016, Theorems 7.16.1–7.16.2). From the point of view of **DTK**, it is in principle impossible to prove or refute either disjunct. Koellner concluded that

> Since the statements that "the mind cannot be mechanized" and "there are absolutely undecidable statements" are independent of the natural principles governing the fundamental concepts and, moreover, are independent of any plausible principles in sight, it seems likely that these statements are themselves "absolutely undecidable". (Koellner, 2018b, p. 469)[27]

[26] For the details of the system **DT** and **DTK**, see (Koellner, 2016; 2018b).

[27] Koellner concluded in (2018b, p. 480) with a disjunctive conclusion of his own: "Either the statements that 'the mind cannot be mechanized' and 'there are absolutely undecidable statements' are indefinite (as the philosophical critique maintains) or they are definite and the above results and considerations provide evidence that they are about as good examples of 'absolutely undecidable' propositions as one might find".

In our previous discussion of GD, the first disjunct and the second disjunct, we identified absolutely undecidability with knowability of the idealized human mind and define that φ is absolutely undecidable if $\varphi \notin \mathbf{K}$ and $\neg\varphi \notin \mathbf{K}$. Under this framework, the second disjunct is equivalent to "**K** is not complete". Under the assumption that $\mathbf{K} \subseteq \mathbf{Truth}$, the second disjunct is equivalent to "$\mathbf{K} \subsetneq \mathbf{Truth}$". However, G1 only tells us that $\mathbf{Prov} \subsetneq \mathbf{Truth}$, and it does not tell us that $\mathbf{K} \subsetneq \mathbf{Truth}$.

Another natural informal definition of absolutely undecidability is: φ is absolutely undecidable if there is no consistent extension T of **ZFC** with well-justified axioms such that φ is provable in T. In this paper, we focus on whether Gödel's incompleteness theorem implies that the human mind cannot be mechanized. In philosophy of set theory, there are extensive discussions about whether there exists an absolutely undecidable statement in set theory. For a detailed discussion of the question of absolutely undecidability in set theory and especially whether the Continuum Hypothesis is absolutely undecidable, we refer to Koellner (2006).

5. Gödel's Undemonstrability of Consistency Thesis and the Definability of Natural Numbers

In Section 8, Krajewski (2020) discussed two consequences of Gödel's incompleteness theorem directly related to the Anti-Mechanist Arguments: Gödel's Undemonstrability of Consistency Thesis and the undefinability of natural numbers. For us, Krajewski's discussion on these two consequences is mainly philosophical and not very precise. In this section, we want to give a more precise logical analysis of Gödel's Undemonstrability of Consistency Thesis and the undefinability of natural numbers.

Let us first examine the definability of natural numbers. As a consequence of Gödel's incompleteness theorem, Krajewski (2020) claimed that we can not define the natural numbers in the sense that there is not a complete axiomatic system which fully characterizes all truths about natural numbers. We give some supplementary notes to make this point more precise.

Firstly, whether a theory about natural numbers is complete depends on the language of the theory. In the languages $L(\mathbf{0}, \mathbf{S})$, $L(\mathbf{0}, \mathbf{S}, <)$ and $L(\mathbf{0}, \mathbf{S}, <, +)$, there are, respectively, recursively axiomatized complete arithmetic theories (see Enderton, 2001, Section 3.1–3.2). For example, Presburger arithmetic is a complete theory of the arithmetic of addition in the language $L(\mathbf{0}, \mathbf{S}, +)$ (see Murawski, 1999, Theorem 3.2.2, p. 222). However, if a recursively axiomatized theory contains enough information about addition and multiplication, then it is incomplete and hence it must miss some truths about arithmetic. For example, any recursively axiomatized consistent extension of **Q** is incomplete. Thus, in Krajewski's sense, we can not define the natural numbers in any recursively axiomatized consistent extension of **Q**.

Secondly, if we discuss the definability of a set with respect to a structure, then the definability of natural numbers depends on the structure we talk about. It is well known that $\mathbb{N}$ is definable in $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{Q}, +, \cdot)$ (Epstein, 2011, Chapter XVI), and $Th(\mathbb{N}, +, \cdot)$ is interpretable in $Th(\mathbb{Z}, +, \cdot)$ and $Th(\mathbb{Q}, +, \cdot)$. Since $Th(\mathbb{N}, +, \cdot)$ is undecidable,[28] by Theorem 2.1, $Th(\mathbb{Z}, +, \cdot)$ and $Th(\mathbb{Q}, +, \cdot)$ are all undecidable and hence not recursive axiomatizable. But $Th(\mathbb{R}, +, \cdot)$ is a decidable, recursively axiomatizable theory (even if not finitely axiomatizable) and $Th(\mathbb{R}, +, \cdot) = \mathbf{RCF}$ (the theory of real closed field; see Epstein, 2011, p. 320–321). As a corollary, $\mathbb{N}$ is not definable in the structure $\langle\mathbb{R}, +, \cdot\rangle$ (if $\mathbb{N}$ is definable in $\langle\mathbb{R}, +, \cdot\rangle$, then $Th(\mathbb{N}, +, \cdot)$ is interpretable in $Th(\mathbb{R}, +, \cdot)$ and thus, by Theorem 2.1, $Th(\mathbb{R}, +, \cdot)$ is undecidable which leads to a contradiction). In summary, if we consider matters of definability relative to the base structure, then whether the set of natural numbers is definable depends on the base structure: $\mathbb{N}$ is definable in $(\mathbb{Z}, +, \cdot)$ and $(\mathbb{Q}, +, \cdot)$, but $\mathbb{N}$ is not definable in $\langle\mathbb{R}, +, \cdot\rangle$.

Now we examine Gödel's Undemonstrability of Consistency Thesis (i.e. G2). The intensionality of Gödel sentence and the consistency sentence has been widely discussed in the literature (e.g. Feferman, 1960; Halbach, Visser, 2014a; 2014b; Visser, 2011). Halbach and Visser examined the sources of intensionality in the construction of self-referential sentences of arithmetic in (2014a; 2014b) and argued 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. Visser (2011) located three sources of indeterminacy in the formalization of a consistency statement for a theory *T*:

(I) the choice of a proof system;
(II) the choice of a way of numbering;
(III) the choice of a specific formula numerating the axiom set 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.

For a consistent theory *T*, we say that G2 holds for *T* if the consistency statement 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. Both mathematically and philosophically, G2 is more problematic than G1. The difference between G1 and G2 is that in the case of G1 we are mainly interested in the fact that it shows that some sentence is undecidable if **PA** is ω-consistent. We make no claim to the effect that that sentence "really" expresses what we would express by saying "**PA**

[28] I.e. there does not exist an effective algorithm such that given any sentence φ in $L(\mathbf{PA})$, we can effectively decide whether $(\mathbb{N}, +, \cdot) \vDash \varphi$ or not.

cannot prove this sentence".[29] But in the case of G2 we are also interested in the content of the statement.

The status of G2 is essentially different from G1 due to the intensionality of G2. 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 the following, we give a very brief discussion of the intensionality of G2 (for more details, we refer to Cheng, in press). In this section, unless otherwise stated, we make the following assumptions:

(1) The theory *T* is a recursively axiomatized consistent extension of **Q**;

(2) The canonical arithmetic formula to express the consistency of *T* is $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0} \urcorner)$;

(3) The canonical numbering we use is Gödel's numbering;

(4) The provability predicate we use is standard;

(5) The formula numerating the axiom set of *T* is Σ^0_1.

Based on works in the literature, we argue that "whether G2 holds for *T*" depends on the following factors:

(1) the choice of the base theory *T*;

(2) the choice of a provability predicate;

(3) the choice of an arithmetic formula to express consistency;

(4) the choice of a numbering;

(5) the choice of a specific formula numerating the axiom set of *T*.

These factors are not independent of each other, and a choice made at an earlier stage may have influences on the choices made at a later stage. In the following, when we discuss how G2 depends on one factor, we always assume that other factors are fixed as in the default assumptions we make and only the factor we are discussing is varied. For example, Visser (2011) rests on fixed choices for (1) and (3)–(5) but varies the choice of (2); Grabmayr (2020) rests on fixed choices for (1)–(2) and (4)–(5) but varies the choice of (3); Feferman (1960) rests on fixed choices for (1)–(4) but varies the choice of (5).

In the following, we give a brief discussion of how G2 depends on the above five factors. For more discussions of these factors, we refer to (Cheng, in press).

"Whether G2 holds for *T*" depends on the choice of the base theory. A foundational question about G2 is: how much of information about arithmetic is re-

[29] I would also like to thank the referee for pointing out this difference between G1 and G2.

quired for the proof of G2. If the base theory does not contain enough information about arithmetic, then G2 may fail in the sense that the consistency statement is provable in the base theory. Willard (2006) explored the generality and boundary-case exceptions of G2 under some base theories. Willard constructed examples of recursively enumerable arithmetical theories that couldn't prove the totality of successor function but could prove their own canonical consistency (see Willard, 2001; 2006). Pakhomov (2019) defined a theory $H_{<\omega}$ and showed 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.

"Whether G2 holds for T" depends on the definition of provability predicate. Recall that T is a recursively axiomatizable consistent extension of **Q**. Being a consistency statement is not an absolute concept but a role w.r.t. a choice of the provability predicate. Note that G2 holds for any standard provability predicate in the sense that if provability predicate $\mathbf{Pr}_T(x)$ is standard, then $T \nvdash \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. However, G2 may fail for some nonstandard provability predicates. Rosser provability predicate is an important kind of non-standard provability predicate in the study of meta-mathematics of arithmetic. Define the Rosser provability predicate $\mathbf{Pr}^R_T(\mathrm{x})$ as the formula $\exists y(\mathbf{Prf}_T(x,y) \wedge \forall z \leq y \neg\mathbf{Prf}_T(\dot{\neg}(x), z))$.[30] Define the consistency statement $\mathbf{Con}^R(T)$ via Rosser provability predicate as $\neg\mathbf{Pr}^R_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. Then G2 fails for Rosser provability predicate: $T \vdash \mathbf{Con}^R(T)$.

"Whether G2 holds for T" depends on the choice of arithmetic formulas to express consistency. We have different ways to express the consistency of T. The canonical arithmetic formula to express the consistency of T is $\mathbf{Con}(T) \triangleq \neg\mathbf{Pr}_T(\ulcorner \mathbf{0} \neq \mathbf{0}\urcorner)$. Another way to express the consistency of T is $\mathbf{Con}^0(T) \triangleq \forall x(\mathbf{Fml}(x) \wedge \mathbf{Pr}_T(x) \rightarrow \neg\mathbf{Pr}_T(\dot{\neg} x))$.[31]

Kurahashi (2019) constructed a Rosser provability predicate such that G2 holds for the consistency statement formulated via $\mathbf{Con}^0(T)$ (i.e. the consistency statement formulated via $\mathbf{Con}^0(T)$ and Rosser provability predicate is not provable in T), but G2 fails for the consistency statement formulated via $\mathbf{Con}(T)$ (i.e. the consistency statement formulated via $\mathbf{Con}(T)$ and Rosser provability predicate is provable in T).

"Whether G2 holds for T" depends on the choice of numberings. 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. Grabmayr (2020) showed that G2 holds for acceptable numberings; But G2 fails for some nonacceptable numberings.[32]

Finally, "Whether G2 holds for T" depends on the numeration of T. As a generalization, G2 holds for any Σ^0_1 numeration of T: if $\alpha(x)$ is a Σ^0_1 numeration of T,

[30] $\dot{\neg}$ is a function symbol expressing a primitive recursive function calculating the code of $\neg\varphi$ from the code of φ.

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

[32] For the definition of acceptable numberings, we refer to (Grabmayr, 2020).

then $T \nvdash \mathbf{Con}_\alpha(T)$. However, G2 fails for some Π^0_1 numerations of T. For example, Feferman (1960) constructed a Π^0_1 numeration $\tau(u)$ of T such that G2 fails under this numeration: $T \vdash \mathbf{Con}_\tau(T)$.

REFERENCES

Beklemishev, L. D. (2010). Gödel Incompleteness Theorems and the Limits of Their Applicability I. *Russian Math Surveys*, *65*(5), 857–899.

Benacerraf, P. (1967). God, the Devil and Gödel. *The Monist*, 51, 9–32.

Boolos, G. (1993). *The Logic of Provability*. Cambridge: Cambridge University Press.

Carlson, T. J. (2000). Knowledge, Machines, and the Consistency of Reinhardt's Strong Mechanistic Thesis. *Annals of Pure and Applied Logic*, *105*(1–3), 51–82.

Chalmers, D. J. (1995). Minds, Machines, and Mathematics: A Review of Shadows of the Mind by Roger Penrose. *Journal Psyche*, *2*.

Cheng, Y. (2019). *Incompleteness for Higher-Order Arithmetic: An Example Based on Harrington's Principle* (Springer series: Springerbrief in Mathematics). New York: Springer.

Cheng, Y. (2020). Finding the Limit of Incompleteness I. *The Bulletin of Symbolic Logic*. Retrieved from: https://arxiv.org/pdf/1902.06658.pdf

Cheng, Y. (in press). Current Research on Gödel's Incompleteness Theorem. *The Bulletin of Symbolic Logic*.

Enderton, H. B. (2001). *A Mathematical Introduction to Logic* (2nd ed.). Boston, MA: Academic Press.

Epstein, R. L. (With contributions by Lesław W. Szczerba). (2011). *Classical Mathematical Logic: The Semantic Foundations of Logic*. Princeton, New Jersey: Princeton University Press.

Feferman, S. (1960). Arithmetization of Metamathematics in a General Setting. *Fundamenta Mathematicae*, *49*, 35–92.

Feferman, S. (1995). Penrose's Gödelian Argument: A Review of Shadows of the Mind by Roger Penrose. *Journal Psyche*, *2*.

Feferman, S. (2009). Gödel, Nagel, Minds, and Machines. *The Journal of Philosophy*, *106*(4), 201–219.

Gaifman, H. (2000). What Gödel's Incompletness Result Does and Does Not Show. *The Journal of Philosophy*, *97*(8), 462–470.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatsh. Math. Phys*. *38*(1), 173–198.

Gödel, K. (1986). An Interpretation of the Intuitionistic Propositional Calculus. In S. Feferman et. al. (Eds.), *Collected Works, Volume I: Publications 1929–1936* (pp. 301–303). Oxford University Press.

Gödel, K. (1951). Some Basic Theorems on the Foundations of Mathematics and Their Implications. In S. Feferman et. al. (Eds.), *Collected Works, Volume III: Unpublished Essays and Lectures* (pp. 304–323). Oxford University Press.

Gödel, K. (1995). *Collected Works, Volume III: Unpublished Essays and Lectures*. New York: Oxford University Press.

Grabmayr, B. (2020). On the Invariance of Gödel's Second Theorem With Regard to Numberings. *The Review of Symbolic Logic*. Retrieved from: https://arxiv.org/pdf/1803.08392.pdf

Hájek, P., Pudlák, P. (1993). *Metamathematics of First-Order Arithmetic*. Berlin-Heidelberg-New York: Springer-Verlag.

Halbach, V., Visser, A. (2014a). Self-Reference in Arithmetic I (2014a). *Review of Symbolic Logic*, *7*(4), 671–691.

Halbach, V., Visser, A. (2014b). Self-Reference in Arithmetic II (2014b). *Review of Symbolic Logic*, *7*(4), 692–712.

Horsten, L. (1998). In Defense of Epistemic Arithmetic. *Synthese*, *116*(1), 1–25.

Horsten, L., Welch, P. (2016). *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge.* Oxford University Press.

Koellner, P. (2006). On the Question of Absolute Undecidability. *Philosophia Mathematica*, *14*(2), 153–188.

Koellner, P. (2016). Gödel's Disjunction. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge*. Oxford University Press.

Koellner, P. (2018a). On the Question of whether the Mind Can Be Mechanized. Part 1: From Gödel to Penrose. *Journal of Philosophy*, *115*(7), 337–360.

Koellner, P. (2018b). On the Question of Whether the Mind Can Be Mechanized. Part 2: Penrose's New Argument. *Journal of Philosophy*, *115*(9), 453–484.

Kotlarski, H. (2004). The Incompleteness Theorems After 70 Years. *Annals of Pure and Applied Logic*, *126*, 125–138.

Krajewski, S. (2020). On the Anti-Mechanist Arguments Based on Gödel Theorem. *Studia Semiotyczne*, *34*(1), 9–56.

Kurahashi, T. (2019). Rosser Provability and the Second Incompleteness Theorem. Retrieved from: https://arxiv.org/pdf/1902.06863.pdf

Leitgeb, H. (2009). On Formal and Informal Provability. In O. Bueno, Ø. Linnebo (Eds.), *New Waves in Philosophy of Mathematics* (pp. 263–299). London: Palgrave Macmillan.

Lindström, P. (1997). *Aspects of Incompleteness*. Cambridge University Press.

Lindström, P. (2001). Penrose's New Argument. *Journal of Philosophical Logic*, *30*, 241–250.

Lindström, P. (2006). Remarks on Penrose's New Argument. *Journal of Philosophical Logic*, *35*, 231–237.

Lucas, J. R. (1961). Minds, Machines, and Gödel. *Philosophy*, *36*, 120–124.

Lucas, J. R. (1996). Minds, Machines, and Gödel: A Retrospect. In P. J. R. Millican, A. Clark (Eds.), *Machines and Thought: The Legacy of Alan Turing* (vol. 1, pp. 103–124). Oxford University Press.

Montague, R. (1963). Syntactical Treatments of Modality, With Corollaries on Reflexion Principles and Finite Axiomatizability. *Acta Philosophica Fennica*, *16*, 153–167.

Murawski, R. (1999). *Recursive Functions and Metamathematics: Problems of Completeness and Decidability, Gödel's Theorems*. Heidelberg: Springer Netherlands.

Myhill, J. (1960). Some Remarks on the Notion of Proof. *Journal of Philosophy*, *57*(14), 461–471.

Nagel, E., Newman, J. R. (2001). *Gödel's Proof*. New York University Press.

Pakhomov, F. (2019). A Weak Set Theory That Proves Its Own Consistency. Retrieved from: https://arxiv.org/pdf/1907.00877.pdf

Pudlák, P. (1999). A Note on Applicability of the Incompleteness Theorem to Human Mind. *Annals of Pure and Applied Logic*, *96*, pp. 335–342.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness*. Oxford University Press.

Penrose, R. (2011). Gödel, the Mind, and the Laws of Physics. In M. Baaz, Ch. H. Papadimitriou, H. Putnam, D. S. Scott, Ch. L. Harper (Eds.), *Kurt Gödel and the Foundations of Mathematics: Horizons of Truth* (pp. 339–358). Cambridge University Press.

Putnam, H. (1960). Minds and Machines. In S. Hood (Ed.), *Dimensions of Mind: A Symposium* (pp. 138–164). New York University Press.

Reid, C. (1996). *Hilbert*. Springer.

Reinhardt, W. N. (1985). Absolute Versions of Incompleteness Theorems. *Noûs*, *19*(3), 317–346.

Reinhardt, W. N. (1985). The Consistency of a Variant of Church's Thesis With an Axiomatic Theory of an Epistemic Notion. *Revista Colombiana de Matematicas*, *19*, 177–200.

Reinhardt, W. N. (1986). Epistemic Theories and the Interpretation of Gödel's Incompleteness Theorems. *Journal of Philosophical Logic*, *15*, 427–474.

Shapiro, S. (1985). Epistemic and Intuitionistic Arithmetic. In S. Shapiro (Ed.), *Intensional Mathematics* (vol. 113 of Studies in Logic and the Foundations of Mathematics, pp. 11–46). Amsterdam: North-Holland.

Shapiro, S. (1998). Incompleteness, Mechanism, and Optimism. *The Bulletin of Symbolic Logic*, *4*(3), 273–302.

Shapiro, S. (2003). Mechanism, Truth, and Penrose's New Argument. *Journal of Philosophical Logic*, 32(1), 19–42.

Smith, P. (2007). *An Introduction to Gödel's Theorems*. Cambridge University Press.

Smoryński, C. (1977). The Incompleteness Theorems. In J. Barwise (Ed.), *Handbook of Mathematical Logic* (pp. 821–865). Amsterdam: North-Holland.

Tarski, A., Mostowski, A., Robinson, R. M. (1953). *Undecidabe Theories* (Studies in Logic and the Foundations of Mathematics). Amsterdam: North-Holland.

Thomason, R. H. (1980). A Note on Syntactical Treatments of Modality. *Synthese*, *44*, 391–395.

Visser, A. (2011). Can We Make the Second Incompleteness Theorem Coordinate Free? *Journal of Logic and Computation*, *21*(4), 543–560.

Visser, A. (2016). The Second Incompleteness Theorem: Reflections and Ruminations. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 67–91). Oxford University Press.

Wang, H. (1996). *A Logical Journey: From Godel to Philosophy*. MIT Press.

Willard, D. E. (2001). Self-Verifying Axiom Systems, the Incompleteness Theorem and Related Reflection Principles. *Journal of Symbolic Logic*, *66*(2), 536–596.

Willard, D. E. (2006). A Generalization of the Second Incompleteness Theorem and Some Exceptions to It. *Ann. Pure Appl. Logic*, *141*(3), 472–496.

Williamson, T. (2016). Absolute Provability and Safe Knowledge of Axioms. In L. Horsten, P. Welch (Eds.), *Gödel's Disjunction: The Scope and Limits of Mathematical Knowledge* (pp. 243–253). Oxford University Press.

Zach, R. (2007). Hilbert's Program Then and Now. In D. Jacquette (Ed.), *Philosophy of Logic* (pp. 411–447). Amsterdam: North Holland.