Russian Math. Surveys **0**:0 1–000 © 1945 RAS(DoM) and LMS
Uspekhi Mat. Nauk **0**:0 61–106 DOI

Gödel incompleteness theorems and the limits of their applicability. I

L. D. Beklemishev

Abstract. This is a survey of results related to the Gödel incompleteness theorems and the limits of their applicability. The first part of the paper discusses Gödel's own formulations along with modern strengthenings of the first incompleteness theorem. Various forms and proofs of this theorem are compared. Incompleteness results related to algorithmic problems and mathematically natural examples of unprovable statements are discussed.

Bibliography: 68 titles.

Keywords: Gödel theorems, incompleteness, proof, computability.

Contents

This work was supported by the Russian Foundation for Basic Research and the programme "Leading Scientific Schools".

AMS 2000 *Mathematics Subject Classification.* Primary 03F40; Secondary 03F25, 03F30, 03F45.

1. Introduction

The Gödel incompleteness theorems are a universally recognized achievement of the mathematical thought of the 20th century. They laid the foundation of mathematical logic. Methods created by Gödel were decisive factors that led to a precise mathematical definition of algorithm, and ultimately to the creation of computers.

At the same time, by their very nature Gödel's theorems touch upon questions going far beyond mathematics proper and connected with topics exciting the imagination, like the mystery of the nature of the human mind, and problems of cognition and artificial intelligence. In this capacity Gödel's theorems, or rather their interpretations, played a significant role in shaping the general intellectual context of the 20th century. As a result, Gödel's theorems are among those mathematical discoveries of the past century which became most widely known outside mathematics itself.[1]

A vast literature has been devoted to Gödel's incompleteness theorems, from quite specialized to pedagogical, popular scientific, and aesthetic. In particular, such worldwide bestsellers as *Gödel, Escher, Bach: the eternal golden braid* by Douglas R. Hofstadter or *The Emperor's new mind* and *Shadows of the mind* by Roger Penrose, belong to the last category. The sea of information devoted to Gödel's theorems mainly contains expositions of these theorems that are accessible to a rather broad circle of readers (primarily of his first theorem, which is simpler). In a sense, the very nature of these results makes it possible to subject Gödel's ideas to boundless variation and adapt them to the taste of any particular author.

In this survey we do not aim to give yet another popular presentation of Gödel's theorems. We would like to fill a gap of another kind, namely, to acquaint the reader with modern generalizations of Gödel's incompleteness theorems — first and foremost, his second theorem — and with diverse and subtle mathematical problems around these results.

Since Gödel's theorems belong to the classics of mathematical logic, we do not go into all the details of proofs which are sufficiently well known in various forms. Instead, we focus on a search for and discussion of optimal formulations of Gödel's results, and also on comparisons of different proofs. In connection with the question of the limits of applicability of Gödel's incompleteness theorems, we present results which have accumulated in mathematical logic, including very recent ones.

The survey divides naturally into two parts, devoted to Gödel's first and second theorems, respectively.

An appropriate context for the first theorem was created by the development of the theory of computable functions, and the role of this theorem became rather understandable in the framework of this theory. However, the second theorem does not lie completely within this context and is more problematic on the whole. Initially, the prevailing opinion was that Gödel's second theorem is a kind of 'supplement' to the first theorem that only indicates an explicit form of a quite independent statement whose existence is asserted in the first theorem. To some extent, this point of view is widespread even now.

[1]Implicit evidence for this assertion is that Kurt Gödel was named, justly or not, as one of Time Magazine's hundred most influential people of the 20th century.

At the same time, after the works of Kreisel, Löb, Feferman, de Jongh, Smoryński, and others who studied generalizations of Gödel's second theorem, it became clear that the nature of these two statements is substantially different. Gödel's second theorem turns out to be mainly connected with modal logical properties of the formula expressing provability, and with the self-reference effect in arithmetics.

In our opinion, it is still too early to decide definitively what the 'correct' context for Gödel's second incompleteness theorem is. At the same time, many partial results have accumulated in this area which clarify the role and the specific features of this theorem. We intend to acquaint the reader with the results and open problems remaining here in the second part of our survey.

In the first part we discuss Gödel's first theorem and its diverse forms. This material is quite traditional, but we decided to include a fairly detailed survey of it for two reasons. First, the problems connected with Gödel's second theorem can be evaluated in full measure only after having formed a very clear idea about various aspects of the first theorem. At the same time, many of the existing presentations of this material are often too one-sided. Second, in the literature one can find many different assertions referred to as a Gödel theorem, not to mention different proofs of these assertions. We would like to systematize this material, at least partially, by considering it from some unified position.

I would like to thank Vladimir Andreevich Uspenskii and Albert Visser for discussions of some of the problems considered here and Sergei Ivanovich Adian for attentive reading and criticism of the manuscript. Their works had also significantly influenced the opinions of the author that are reflected in this survey.

2. Gödel's first and second incompleteness theorems

In the fundamental paper [1], Gödel proved his theorems 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. At that time, *Principia Mathematica* was perhaps the most well-known system of axioms intended for the formalization of mathematics. Gödel also explained that his result can be extended to other axiomatic systems, including Zermelo–Fraenkel set theory, von Neumann set theory, and other theories "developed recently by D. Hilbert and his disciples." We do not present a precise definition of the system P, the formulation of which is somewhat cumbersome. Stronger results will be obtained for the language of first-order arithmetic which is standard nowadays.

The simplest formulation of Gödel's first incompleteness theorem asserts that there is a sentence which is neither provable nor refutable in the theory P under consideration. Gödel's second incompleteness theorem asserts that for this sentence one can take a formalization in P of the statement that the theory P itself is consistent.

The incompleteness of theories like P (or set theories especially created for the axiomatization of the whole of mathematics) drastically contradicted the opinions prevailing at that time. Moreover, Gödel's second theorem placed in doubt the possibility of realizing the most important thesis of the so-called *Hilbert programme*

(see [2]), whose proclaimed objective was to establish the consistency of mathematics (analysis and set theory) by using finitary tools. This problem was regarded by the representatives of Hilbert's school as the central problem of mathematical logic. However, it follows from Gödel's second theorem that it is impossible to formalize the 'finitary tools' that are able to establish the consistency of mathematics even in the framework of a very strong system P.[2]

In principle, the incompleteness of some specific theory, say, of P, can mean only that some necessary axioms "were not taken into account." (For example, this was the case for a long time with the axiomatization of elementary geometry.) To show that, in the present case, we face a substantially different and more dramatic situation, Gödel formulates his first theorem in a more general form, which speaks of the *fundamental incompletability* of the system P.

Keeping the meaning sufficiently close to the original formulation, we can state *Gödel's first incompleteness theorem* ([1], Theorem VI) as follows.

Theorem 1. *Let T be a formal theory satisfying the following conditions*:

(i) *T is formulated in the language of P*;
(ii) *T is obtained by adding a primitive recursive set of axioms to the system P*;
(iii) *T is ω-consistent.*

Then T is incomplete, that is, there is a sentence φ for which neither φ nor $\neg\varphi$ is provable in T.

Let us clarify the notions used here and the conditions (i)–(iii) themselves. It should be noted immediately that none of these conditions is the most general or the most natural mathematically from the modern point of view. (The reasons for this are briefly discussed below; see also the corresponding remarks in [3] and [4].) However, these conditions are of interest: they reflect the initial perception by Gödel of his discovery in 1931, and moreover, the use of these conditions was to a significant degree fixed in the subsequent literature.

The condition (ii) involves the notion of primitive recursive function.

Definition 1. A function $f\colon \mathbb{N}^k \to \mathbb{N}$ is said to be *primitive recursive* if it can be obtained from the constant 0, the successor function $S(x) = x + 1$, and the projection functions $I_n^m(x_1, \dots, x_n) = x_m$ by using the composition operation (substitution) and primitive recursion. A function f is said to be *obtained from g and h by primitive recursion* if

$$\begin{cases} f(0, \vec{x}\,) = g(\vec{x}\,), \\ f(n+1, \vec{x}\,) = h\big(f(n, \vec{x}\,), n, \vec{x}\,\big). \end{cases}$$

A set $R \subseteq \mathbb{N}^k$ is said to be *primitive recursive* if the assertion $\vec{x} \in R$ is equivalent to $f(\vec{x}\,) = 0$ for some primitive recursive function f. (As is customary in logic, we shall represent the assertion that $\vec{x} \in R$ also in the form $R(\vec{x}\,)$.)

[2] Gödel expresses this fact in a curious way, apparently trying to avoid a philosophical controversy: "It should be especially stressed that Theorem IX [the second incompleteness theorem] does not contradict Hilbert's formalistic point of view. Indeed, [...] one can imagine that there are finitary proofs which *cannot be* formalized in P." Nowadays, it is customary to think that finitary proofs can be formalized even in a much weaker first-order Peano arithmetic. At the same time, there are alternative opinions concerning this question, originating from Hilbert himself (see [2]).

The condition (ii) presupposes that some one-to-one encoding of the objects of the language of the theory P by the natural numbers (a Gödel numbering) is fixed in advance, where the objects include variables, terms, formulae, and so on. Gödel proved that for some natural choice of this encoding the relation "x is the code of the derivation of a formula with the code y in P" is primitive recursive.

The condition (ii) means that the set of codes of the axioms of T for a chosen encoding must be primitive recursive. Since T is constructed on the basis of P and the deductive mechanism of P is primitive recursive, this implies that the relation "x is the code of the derivation of a formula y in T" is also primitive recursive. The assumption (ii) is rather general and holds for all formal theories considered in practice. Gödel noted that this condition certainly holds for all theories presented by finitely many axioms or schemes of axioms. At present, it is customary to refer to the theories satisfying the condition (ii) as *primitively recursively axiomatized* theories.

Definition 2. A theory T is said to be *ω-consistent* if there is no formula $\varphi(x)$ (where the variable x ranges over the natural numbers) such that the following conditions hold simultaneously:

(i) $T \vdash \exists x\, \varphi(x)$;
(ii) $T \vdash \neg\varphi(\underline{0}), \neg\varphi(\underline{1}), \dots$.

We recall that for the theories T in the language of P every natural number n is represented by the closed term $S(S(\dots S(0)\dots))$ (n times), which we denote by $\underline{n}$ and refer to as a *numeral*. We abbreviate a sequence of numerals $(\underline{n}_1, \dots, \underline{n}_k)$ by $\vec{\underline{n}}$.

The condition of ω-consistency strengthens the consistency condition for a theory T. In turn, this condition follows from the assumption that T is *sound*, that is, all theorems of T are true in the model with the support $\mathbb{N}$. However, Gödel himself did not consider the semantic concept of soundness in his paper.

The condition (iii) is not essentially restrictive either, namely, the theories which do *not* satisfy this condition can be regarded as theories close to being contradictory, that is, as a pathological case. In fact, Gödel says that if the natural and broad assumptions (i) and (ii) are satisfied, then we face a choice between two 'unpleasant' possibilities: either the theory T is incomplete or it is ω-contradictory.

As for the condition (i), it is rather restrictive. Natural theories certainly need not be formulated literally in the language of P, for example, this is so for set theory. In this connection Gödel notes in his comments to Theorem VI that, under the conditions of the theorem, it suffices to assume instead of (i) that all primitive recursive relations are *decidable*[3] (entscheidungsdefinit) in the theory T, and instead of (ii) that the set of codes of the axioms of T is decidable in T.

Definition 3. A relation R is said to be *decidable in* T if there is a formula $\varphi_R(\vec{x})$ such that

$$R(\vec{n}) \Rightarrow T \vdash \varphi_R(\vec{\underline{n}}), \tag{1}$$
$$\neg R(\vec{n}) \Rightarrow T \vdash \neg\varphi_R(\vec{\underline{n}}) \tag{2}$$

for any tuple $\vec{n}$ of numbers.

[3]Different sources refer to this notion also as *binumerability*, *representability*, or *definability in the theory*.

The central point of Gödel's proof was the theorem on the decidability in P of all primitive recursive relations (however, Gödel proves this theorem only schematically).[4] In particular, this enables him to express an independent statement for the theory T in the form $\forall x\, \varphi_R(x)$, where R is primitive recursive. The additional idea of encoding finite number sequences by using the so-called *Gödel β-function* (based on the Chinese remainder theorem) shows ([1], Theorem VII) that every primitive recursive relation is equivalent (in P) to some arithmetical formula, that is, a formula of a first-order language in the signature with the constant 0, the functions S, $+$, and $\cdot$, and the equality relation. Thus, Gödel proved that the independent statement constructed by him for the theory T relates to elementary arithmetic, whatever the language of the theory T itself is.

Discussion. To appreciate the conditions of Theorem 1 from the modern point of view, one must take into account the following facts.

First, the theory of computability, which gives the correct context for Gödel's first theorem, had not yet been created at that time. In fact, it was Gödel's theorems that to a great extent stimulated the creation of this theory: the notion of general recursive function was introduced in Gödel's 1934 lectures [5], and the equivalence of this notion to the notion of an (everywhere defined) computable function was later proved in other formulations by Church, Turing, and Kleene. Gödel himself noted, in an addendum to the English translation of his paper in 1963, the important role of Turing, thanks to whom "a completely general version of Theorems VI and XI is now possible" (that is, of Gödel's first and second theorems; see [6], p. 195).

Second, Gödel avoids using semantic notions, in particular, the notion of model and the notions of truth and definability in a model. This is apparently related to two circumstances. First of all, before Gödel's paper, it was not quite clear what the difference is between the notions of provability and truth for theories like P. Moreover, the semantic notions in logic on the whole were under suspicion due to paradoxes well known to everyone. The important paper [7] of Tarski, which deals with the investigation of the notion of truth for formalized languages (and which significantly reduced these suspicions), was published only in 1933 (and in German in 1935). This can also explain the occurrence of the ω-consistency condition instead of the more fundamental condition of semantic soundness. There was also no rigorous notion of interpretation of one theory in another (this notion was formed much later under the influence of Tarski's works), and this deficiency led, in particular, to restriction to theories in the language of P in the formulation of the incompleteness theorems.

On the other hand, Gödel's paper was written mainly counting on readers sharing Hilbert's opinion on the foundations of mathematics. This fact also influenced the style and the choice of some formulations, to the detriment of more natural semantic assertions. Gödel tried to show that the incompleteness theorems themselves belong to finite mathematics and have no connection with anything irrational, with the 'illegal use of infinity' or anything similar.

[4] A rigorous proof of this theorem was presented in the lectures given in 1934 for a certain second-order theory which is related to the system P [5]. It is of interest that, thanks to the presence of set variables in the language of P, the proof of this theorem for P is significantly simpler than for the language of first-order arithmetic, which is customary at present.

Third, in Gödel's paper there is also no clear distinction between the notions of theory and metatheory. Although the terms 'metalanguage' and 'metatheory' go back to Hilbert, the fundamental property of these notions for the circle of problems under consideration became clear thanks to Gödel's paper and to Tarski's paper mentioned above.

Apparently, a correct reading of Gödel's paper consists in an implicit choice of the system P itself as a metatheory in which all meaningful statements in the paper are formalizable. This is shown by a special formalized style of the formulations of all statements in the paper (and a special font is used for the formal analogues of meaningful syntactic notions), beginning with Theorem V on the decidability of primitive recursive relations in P.[5] Gödel himself notes in describing the scheme of the proof of his second theorem that a significant part of the statements in the paper are formalizable in P. Possibly, the choice of such a formal style of presentation was dictated by the desire to help the reader believe that such a formalization is possible.

Fourth, Gödel's notion of decidability in a theory plays the role of a suitable substitute for the semantic notion of definability. It stresses that this notion does not appeal to the 'contentual' meaning (inhaltliche Deutung) of the formulae of the system P. However, one can still see that this notion implicitly appeals to a semantic interpretation of primitive recursive schemes, because the formula φ_R is in fact constructed from a primitive recursive scheme defining R rather than from the semantic relation expressed by R. It should be noted that Gödel identifies primitive recursive functions and schemes, sometimes in a not quite correct way, for example, when defining the *level* of a primitive recursive function (see also Kleene's comments about this). From a philosophical point of view, the semantic interpretation of primitive recursive functions is simpler and more clear in a sense than that for arbitrary formulae of the higher-order language of P.

In connection with the notion of decidability in the theory, another important circumstance must be noted. Later analysis of this notion (see [8]) showed that for the consistent primitively recursively axiomatized theories containing P (and even the much weaker Robinson arithmetic Q), this notion is coextensive with the notion of algorithmic decidability. Thus, a more general formulation of Gödel's first theorem (given by him as a comment on p. 190 of [1] [6]) belongs in essence to the same class of theories as the modern formulation considered in the next section. Here let us consider a formulation of this comment more explicit than that used by Gödel himself, for comparison with subsequent statements.

Statement 1 (Gödel's comment). *Let T be a formal theory satisfying the following conditions*:

(i) *all primitive recursive relations are decidable in T*;

(ii) *the set of axioms and the set of inference rules for the system T (that is, the immediate consequence relation in T) are primitive recursive or at least decidable in T*;

(iii) *T is ω-consistent.*

[5] In his deep commentary [4] to Gödel's paper, Kleene also notes this fact but does not explain it, and refers to it simply as "Gödel's propensity to speak in terms of his numbers."

[6] The preceding remark by Gödel on p. 189 generalizes the condition of primitive recursive axiomatizability to the decidability in T of the set of axioms of the theory T.

Then T is incomplete.

Gödel does not specify here the language of the theories under consideration. However, all three conditions, including (i), are formulated in terms of an abstract 'immediate consequence' relation and the corresponding notion of derivability. This is very close to the abstract notion of a formal system treated as an algorithm generating theorems of T originating from the axioms by the inference rules, the set of which is primitive recursive (or at least decidable). Therefore, the choice of a language does not play a great role in this case, provided that the notion of decidability makes sense for the theory at hand.[7] Of course, this notion was not used by Gödel explicitly; recall that Gödel stressed the role of Turing and others in the subsequent development of this notion.

We also note that Statement 1 can apply also to theories which are not enumerable. For example, for T one can take a first-order theory axiomatized over P by all formula $\forall x\, \varphi(x)$ such that $\varphi(\underline{n})$ is provable in P for any specific $n \in \mathbb{N}$ (a single application of Carnap's ω-rule; see [10]). Gödel himself did not consider such theories in his paper. Nevertheless, they have been playing an important role in investigations of the foundations of mathematics.

3. Modern formulations of Gödel's first theorem

As a result of the emergence of a general notion of computable function in the works of Gödel, Church, Turing, Post, and Kleene, it became clear that the class of primitive recursive functions is less fundamental than the class of arbitrary computable functions or the class of computably enumerable (c.e.) relations. Correspondingly, it is natural to generalize the condition (ii) of primitive recursive axiomatizability of a theory T in Gödel's first theorem to the condition that the set of theorems of T is c.e. This generalization extends the class of axiomatizations under consideration but does not influence the class of theories under consideration, because, as noted later by Craig [11], every c.e. first-order theory can be defined by a primitive recursive set of axioms.

Apparently, Kleene [12] was one of the first to comprehend Gödel's results from the point of view of a more general computability theory. In fact, this early paper of Kleene already contains the main ideas of the approach to Gödel's theorems based on the recursion theory developed by Kleene, ideas which lay at the base of practically all subsequent investigations around Gödel's theorems. In the same paper Kleene gave a general recursion-theoretic version of the first incompleteness theorem (in the semantic variant). From the point of view of the evolution of the formulations, Kleene's formulation occupies an intermediate position between Gödel's informal comments to his theorems and the more perfect abstract formulations which can be found both in later works of Kleene himself [13]–[15] and in works of others, for example, Smullyan [16] and Uspenskii [17].

[7] For first-order languages (or higher-order languages) containing names for natural numbers the notion of decidability given by Gödel is quite rigorous. In the general case one can formulate rather simple abstract conditions on the language for the notion of decidability to make sense in the corresponding formal system. This leads, for example, to Smullyan's notion of *representation system* [9].

On the other hand, Rosser [18] modified Gödel's construction in such a way as to establish the incompleteness of a theory T already under the assumption of its ordinary consistency. This beautiful and stronger form of Gödel's first theorem is sometimes referred to as Rosser's theorem or the Gödel–Rosser theorem. (Nevertheless, Gödel's original construction maintains its value in connection with the proof of the second incompleteness theorem.)

The subsequent technical improvements of the first incompleteness theorem are connected with a series of results of Kleene, Mostowski, Tarski, and R. Robinson. In particular, using Gödel's work, Kleene [13] and Mostowski [19] studied the relations defined in the language of first-order arithmetic and introduced a hierarchy of arithmetical classes Σ_n and Π_n.

Tarski initiated a systematic investigation of decidability problems for first-order theories. In the framework of this programme, quite broad generalizations of Gödel's results were obtained in Tarski's school. On the one hand, very weak arithmetical theories were found, like Robinson's arithmetic Q, for which essential undecidability was established. The Gödel–Rosser theorem was thereby extended to a broader class of theories. On the other hand, the method of interpretation developed by Tarski enabled one to extend these results to other languages different from the arithmetical one. We note that one of the earliest and most beautiful results in this direction was Quine's theorem [20] establishing the mutual definability of the elementary arithmetic of the natural numbers and the theory of concatenation of binary words. In combination with the interpretation method, the Gödel–Rosser theorem was one of the main tools of the proof of the algorithmic undecidability of quite diverse theories.

After the publication of the fundamental monographs of Kleene [15] and of Tarski, Mostowski, and Robinson [21] in the early 1950s, the presentations of Gödel's incompleteness theorems acquired their standard forms, which are very close to the modern ones.

3.1. General formulations. Let us first give very general formulations of syntactic versions of Gödel's theorem and the Gödel–Rosser theorem for abstract formal systems. These formulations have the advantage that they are applicable in many situations and do not depend on the language of a theory. Moreover, they clarify the computational essence of Gödel's first theorem.

We begin with the definition of an abstract formal system.

Definition 4. By a *formal system* we mean a triple $S = (L, P, R)$, where $L \subseteq \mathbb{N}$ is the decidable set of all *sentences* of the system, and $P, R \subseteq L$ are c.e. sets of *provable* and *refutable* sentences, respectively. We assume that all three sets are given by fixed algorithms (Turing machines) denoted by M_L, M_P, and M_R.

We note that we have identified the formal objects of the system and their Gödel numbers, and therefore the corresponding algorithms work on the natural numbers. A system S is said to be *consistent* if $P \cap R = \varnothing$ and *complete* if it is consistent and $P \cup R = L$.

We note that if a formal system S is complete, then the sets P and R are decidable (Post's theorem). Therefore, to prove the incompleteness of S, it suffices to establish the algorithmic undecidability of either of these two sets. The system S is said to be *decidable* if P is.

The undecidability of a system S follows from the existence on a c.e. undecidable set and from the natural assumption that the system S 'expresses' in a sense all c.e. sets. A quite general notion of expressibility can be defined for abstract formal systems as follows.

Definition 5. A set $A \subseteq \mathbb{N}$ is said to be *expressible in* $S = (L, P, R)$ if there is an everywhere defined computable function f such that

$$n \in A \iff f(n) \in P.$$

A system S is said to be *universal for c.e. sets* if all c.e. sets are expressible in it.

The notion of expressibility for formal systems is simply the notion, familiar in the theory of algorithms, of m-reducibility to the set of provable sentences of S. Every set expressible in a decidable system must be algorithmically decidable. Thus, universal formal systems are undecidable and incomplete. This fact can be viewed as an abstract version of Gödel's first incompleteness theorem.

Theorem 2. *If S is a formal system universal for c.e. sets, then S is undecidable and incomplete.*

The Gödel–Rosser theorem is based on another fact of the theory of computable functions, namely, the existence of an inseparable pair of disjoint c.e. sets. We note that to any ordered pair of disjoint sets $A, B \subseteq \mathbb{N}$ one can assign a partial function $g\colon \mathbb{N} \to \{0, 1\}$ for which

$$g(n) = \begin{cases} 0 & \text{if } n \in A, \\ 1 & \text{if } n \in B, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

This correspondence between the pairs and the functions is one-to-one. Both sets A and B are c.e. if and only if the function g is computable.

A system $T = (L_T, P_T, R_T)$ is said to be an *extension of* $S = (L_S, P_S, R_S)$ if $L_T \supseteq L_S$, $P_T \supseteq P_S$, and $R_T \supseteq R_S$. In this case the function g_T corresponding to the pair (P_T, R_T) is an extension of the function g_S.

A pair (A, B) is said to be (*recursively*) *inseparable* if the function g has no everywhere defined computable extension $g'\colon \mathbb{N} \to \{0, 1\}$. A formal system $S = (L, P, R)$ is said to be *inseparable* if the pair (P, R) is.

It follows from what was said above that inseparable formal systems are undecidable and incomplete, and so are any consistent extensions of these systems. Thus, for a formal system inseparability is another sufficient condition for incompleteness.

Definition 6. A pair (A, B) is *separable in a system* $S = (L, P, R)$ if there is a computable function f such that

$$\begin{cases} n \in A \Rightarrow f(n) \in P, \\ n \in B \Rightarrow f(n) \in R. \end{cases}$$

We say that S *separates pairs of c.e. sets* if every pair of disjoint c.e. sets (A, B) is separable in the system S.

Let $\varphi_e(x)$ be the output of a Turing machine with the index e on the input x. It is easy to see that the function

$$F(x) = \begin{cases} 1 & \text{if } \varphi_x(x) = 0, \\ 0 & \text{if } \varphi_x(x) \text{ is defined and } \varphi_x(x) \neq 0, \\ \text{undefined} & \text{otherwise} \end{cases} \tag{3}$$

is computable. However, it cannot have any everywhere defined computable extension. Indeed, suppose that a computable function $g = \varphi_n$ is everywhere defined. Consider $m = g(n) = \varphi_n(n)$. If $m = 0$, then $F(n) = 1$, and if $m \neq 0$, then $F(n) = 0$. In any case $F(n) \neq m = g(n)$, that is, g does not extend F. Thus, inseparable pairs of c.e. sets exist.

This implies the following abstract version of the Gödel–Rosser theorem.

Theorem 3. *Let a formal system $S = (L, P, R)$ be consistent and let it separate pairs of c.e. sets. Then S is inseparable, and thus every consistent extension of S is undecidable and incomplete.*

This statement, as well as the notion of inseparable pair of c.e. sets, goes back to Kleene's paper [14], where it was called *the symmetric form of Gödel's theorem.*

Remark 1. We note that Theorems 2 and 3 are very abstract, and their application to specific formal systems arising in logic needs additional work (some examples are considered below). Therefore, one should not literally identify these statements with Gödel's and Rosser's theorems.

Remark 2. There are examples of non-enumerable theories for which even the abstract formulations given above are not sufficiently general. Diverse generalizations of these theorems, in terms of the so-called *representation systems*, were considered in great detail by Smullyan [9]. Since the central point for us is Gödel's second theorem, we do not dwell on these results.

3.2. Languages, theories, and interpretations.

Logical languages. Applications of the abstract results in § 3.1 relate mainly to formal systems arising on the basis of logical languages, and in particular, on the basis of the language of first-order predicate logic.

We consider first-order languages and theories with equality. In essence, this choice is not restrictive, because, as is well known, one can interpret richer languages in languages of this kind by extending the signature.

First of all, one can interpret a many-sorted language by including in the signature one-place predicate symbols selecting new sorts of variables. Further, introducing a new sort of variables for the functions f and an additional operation $\mathrm{ap}(f, x)$ whose value is the result $f(x)$ of applying a function f to an argument x, we interpret a second-order language with function variables. We can similarly interpret predicate variables and also higher-order variables. Thus, the results presented below can be applied to many-sorted theories and to higher-order theories. This covers a significant part of those logical theories intended for the formalization of mathematics that are considered in practice.

Sometimes fragments of a first-order language are considered for which versions of Gödel's theorems also hold. In particular, languages with bounded quantifiers, quantifier-free languages, and equational languages are of interest from this point of view. The interest in such poor languages is connected first of all with the question of how 'simple' Gödel's independent assertion can be. On the other hand, there is traditional philosophical and historical interest in the subject of Hilbert's programme and in the formalization of 'finite mathematics'. In this context, quantifier-free calculi play an important role.

Theories. By *theories* S we mean first-order theories with equality, that is, theories defined by some signature Σ and a set $\mathscr{P}$ of formulae of this signature which is closed under the logical successor (in the first-order predicate logic with equality). The formulae of $\mathscr{P}$ are said to be *provable* in S.

A theory $S = (\Sigma, \mathscr{P})$ is said to be *computably enumerable* (c.e.) if the signature Σ is decidable and the set $\mathscr{P}$ is c.e. Since Σ is decidable, so is the set $\mathscr{L}_\Sigma$ of all formulae of the given language. When speaking about a c.e. theory S, we assume that some Turing machines defining Σ and $\mathscr{P}$ are fixed.

As a rule, the Turing machine enumerating $\mathscr{P}$ is defined by the deductive mechanism of the theory S. For example, if S is given by a finite family of axioms and inference rules on the basis of Hilbert's predicate calculus, then corresponding to this definition is a certain algorithm enumerating all possible derivations (and derivable formulae) in S. However, in principle the mechanism enumerating the theorems of S can be quite different. For example, S can appeal to non-deductive decision procedures to clarify the truth value of various special classes of formulae (say, by using an algorithm for solving equations if this turns out to be necessary in the process of seeking a conclusion). By the way, mixed algorithms of this kind are typical for computer systems now under intensive development for seeking conclusions.

It is also well known that Hilbert's standard derivation format is inconvenient in many respects for practical work with formal derivations. In proof theory there are developments of diverse alternative deductive systems which are convenient for various applications, for example, Gentzen sequent systems, so-called natural deduction systems, and others (see, for example, [22]). For Gödel's first theorem, the choice of a specific deductive mechanism does not play a large role.

By a *Gödel numbering* of formulae of the language of $\mathscr{L}_\Sigma$ we mean a one-to-one and computable (in both directions) correspondence between the set $\mathscr{L}_\Sigma$ of formulae and some decidable subset $L \subseteq \mathbb{N}$. We note that in choosing some Gödel numbering of the formulae of the language of $\mathscr{L}_\Sigma$, we assign some formal system to the c.e. theory S. (A formula φ is assumed to be refutable in S if $\neg\varphi$ is provable in S. Since the operation $\varphi \mapsto \neg\varphi$ is computable, it follows that the set of refutable formulae of a c.e. theory is also c.e.)

Formal arithmetic. Peano arithmetic PA is a first-order theory with equality in the language containing the constant 0 and symbols for the successor function $S(x) = x + 1$, the addition $+$, and the multiplication $\cdot$. The standard model of PA is the set $\mathbb{N}$ of natural numbers (with zero), considered together with all these operations. The axioms of PA, along with the logical axioms and the equality axioms, are

P1. $\neg S(x) = 0,\ S(x) = S(y) \to x = y,$
P2. $x + 0 = x,\ x + S(y) = S(x + y),$
P3. $x \cdot 0 = 0,\ x \cdot S(y) = (x \cdot y) + x,$
together with the *scheme of induction axioms*

$$\varphi(0) \wedge \forall x \left(\varphi(x) \to \varphi(S(x))\right) \to \forall x\, \varphi(x)$$

for all arithmetical formulae $\varphi(x)$ (possibly containing parameters, that is, free variables besides x).

The *Robinson arithmetic* Q is obtained by replacing the induction scheme in the formulation of PA by the following axiom (obviously derivable in PA by induction on x):

$$x = 0 \vee \exists y\ y = S(x).$$

The system Q traditionally plays a large role in strengthenings of Gödel's theorems.

Interpretations. The notion of relative interpretation is well known for first-order languages and theories (see, for example, [23]). Apparently, interpretations became widely used in mathematical logic after Tarski's works and after the publication by Tarski, Mostowski, and Robinson of the monograph [21], where interpretations were actively used to study decidability problems of logical theories.

Gödel's theorems hold for languages and theories that are universal in a certain sense. However, to speak of universality, we must be able to compare languages, with one another, that is, to deal with some notion of interpretation. From this point of view, sufficiently general formulations of Gödel's theorems *assume* the use of interpretation in a natural way.[8]

We shall use a rather broad notion of interpretation, which can be described as the notion of multidimensional interpretation with definable parameters and non-absolute equality relation (for the precise definition, see § 9). Interpretations of this kind will simply be referred to as *interpretations*. The translation of a formula φ under an interpretation I is denoted by φ^I.

When we fix some interpretation of the signature Σ_1 in Σ_2, we actually choose in the language of Σ_2 a new sort of variables that corresponds to objects of the language of Σ_1, and also the predicates and functions from Σ_1 over these objects (the expressive capabilities of the language of Σ_2 remain the same here). Therefore, we often speak of languages and theories as if these are many-sorted.

One of the central notions for our purposes is the notion of *arithmetical theory*. By an arithmetical theory we mean a theory in which an interpretation of the predicate calculus in the signature of PA is fixed. Informally, we may assume that a special sort of objects is distinguished in the language of an arithmetical theory, namely, the natural numbers, together with the usual arithmetical operations over these objects. However, we do not assume *a priori* that any non-logical axioms for these symbols are provable.

Remark 3. Anticipating, we note that in principle one can transform a theory into an arithmetical theory in many ways, and this can influence the metamathematical

[8] Nevertheless, in most presentations of Gödel's theorems this fact remains latent or is mentioned only on an informal level.

properties of the resulting theory. Consider two standard formalizations of set theory: the Gödel–Bernays formalization GB, and the Zermelo–Fraenkel formalization ZF_g, where ZF_g is considered in the Gentzen sequent format without the cut-rule. As is well known, one can choose an interpretation I of Robinson's arithmetic Q in the set theory GB in such a way that the consistency assertion for ZF_g holds for this interpretation, that is, $\mathsf{GB} \vdash \mathrm{Con}(\mathsf{ZF}_g)^I$ (about this see, for example, [24]). On the other hand, this is not the case for the ordinary von Neumann interpretation J of the natural numbers in set theory. Indeed, GB conservatively extends ZF_g, and this fact can be proved in Peano arithmetic PA. However, the axioms of PA are valid under the von Neumann interpretation of the natural numbers in GB, that is, $\mathsf{GB} \vdash \mathsf{PA}^J$. Thus, by Gödel's second theorem, $\mathsf{GB} \nvdash \mathrm{Con}(\mathsf{ZF}_g)^J$.

3.3. Σ_1-definability. In the language of arithmetic the inequality $x \leqslant y$ is usually expressed as $\exists z\,(z + x = y)$. Let us add the predicate symbol $\leqslant$ to the signature of arithmetic and assume that the equivalence

$$x \leqslant y \quad \longleftrightarrow \quad \exists z\,(z + x = y)$$

is another axiom of Q. In the language thus extended, one introduces as abbreviations the following *bounded quantifiers*:

$$\forall x \leqslant t\,\varphi \quad \stackrel{\text{def}}{\Longleftrightarrow} \quad \forall x\,(x \leqslant t \to \varphi),$$
$$\exists x \leqslant t\,\varphi \quad \stackrel{\text{def}}{\Longleftrightarrow} \quad \exists x\,(x \leqslant t \wedge \varphi),$$

where the term t does not contain the variable x. A formula φ is said to be *bounded* if every occurrence of a quantifier in φ is bounded, that is, has the form $\forall x \leqslant t\,\psi$ or $\exists x \leqslant t\,\psi$. The set of all bounded formulae is denoted by Δ_0.

The classes of Σ_n- and Π_n-formulae are defined by induction on n as follows. We regard the bounded formulae as both Σ_0- and Π_0-formulae. The Σ_{n+1}-formulae are those of the form $\exists \vec{x}\,\varphi(\vec{x}, \vec{y}\,)$, where φ is a Π_n-formula. The Π_{n+1}-formulae are those of the form $\forall \vec{x}\,\varphi(\vec{x}, \vec{y}\,)$, where φ is a Σ_n-formula.

The following fundamental theorem speaks of the coincidence of the classes of c.e. and Σ_1-definable sets in the standard model of $\mathbb{N}$.

Theorem 4. *A set $A \subseteq \mathbb{N}^k$ is c.e. if and only if*

$$\vec{n} \in A \quad \Longleftrightarrow \quad \mathbb{N} \vDash \varphi[\vec{n}\,]$$

for some Σ_1-formula $\varphi(\vec{x}\,)$. Here the formula φ can be constructed effectively (in polynomial time) from the Turing machine defining the set A, and conversely.

The essence of this theorem is that the language of arithmetical Σ_1-formulae which can be interpreted in $\mathbb{N}$ is a universal model of computations. In the sense of this model a Σ_1-formula defining the graph of a partial function can be regarded as a program for computing this function, a program that can be constructed effectively from the Turing machine defining the function.

Corollary 1. *The set of all true Σ_1-sentences is c.e. and undecidable. The set of all true Π_1-sentences is not c.e.*

Proof. The second part follows from the first by Post's theorem, because the true Π_1-sentences effectively correspond to false Σ_1-sentences. Let us prove the first statement.

The enumerability of the set $\mathrm{Th}_{\Sigma_1}(\mathbb{N})$ of true Σ_1-sentences follows from the fact that there is an obvious algorithm for verifying the truth of a given Δ_0-formula $\varphi(\vec{x}\,)$ on a given tuple $\vec{n}$ of arguments. An algorithm accepting exactly the true formulae of the form $\exists\vec{x}\,\varphi(\vec{x}\,)$ for $\varphi \in \Delta_0$ exhausts all possible arguments $\vec{n}$ until the first tuple validating φ is found.

Now let $K \subseteq \mathbb{N}$ be a c.e. undecidable set. Consider a Σ_1-formula φ_K defining K in $\mathbb{N}$. For any $n \in \mathbb{N}$ we have

$$n \in K \iff \mathbb{N} \vDash \varphi_K[n] \iff \mathbb{N} \vDash \varphi_K(\underline{n}\,).$$

We note that the Σ_1-formula $\varphi_K(\underline{n}\,)$ can be constructed from n effectively. Therefore, the question of whether or not n is in K reduces to the question of whether or not the Σ_1-sentence $\varphi_K(\underline{n}\,)$ is true. Hence, the last question cannot be solved algorithmically. This proves Corollary 1.

Gödel's result on the arithmeticity of the primitive recursive relations was a predecessor of Theorem 4. In fact, Gödel indicated a way to construct, from a given primitive recursive scheme, a certain arithmetical formula which could be called a *generalized* Σ_1*-formula*. Feferman [25] refers to these as RE-formulae. The formulae obtained by Gödel's construction admit unbounded existential quantifiers in the scope of a bounded universal quantifier. However, it could be noted that the *scheme of* Σ_1*-boundedness*

$$\forall x \leqslant y\, \exists z\, \varphi(x,y,z) \quad\longleftrightarrow\quad \exists u\, \forall x \leqslant y\, \exists z \leqslant u\, \varphi(x,y,z)$$

holds in the standard model of arithmetic, and this enables one to transform a generalized Σ_1-formula to a Σ_1-formula. Thus, Theorem 4 easily follows from Gödel's original construction.

The classes of arithmetical Σ_n- and Π_n-predicates arose in papers of Kleene [13] and Mostowski [19].

The notion of Δ_0-formula and of the corresponding class of predicates on $\mathbb{N}$ was introduced in the book [16] of Smullyan. Interesting independent characterizations were later obtained for this class, in particular, in terms of the theory of complexity of computations. At present, the class Δ_0 plays an important role in investigations concerning bounded arithmetic (see, for example, [26]). The application of Δ_0-definable and Σ_1-definable relations instead of primitive recursive ones in the proof of Gödel's theorem makes it possible to avoid the unnecessary (in essence) formalism of primitive recursive schemes and to work directly in the language of arithmetic (see [16]).

Strengthenings of Theorem 4 connected with possible additional restriction of the class of Σ_1-formulae are well known. For example, the well-known Matiyasevich theorem, which was based on previous results of M. Davis, H. Putnam, and J. Robinson and gave a negative solution of Hilbert's tenth problem, asserts that every c.e. relation is definable even by a formula of the form $\exists\vec{y}\,A(\vec{x},\vec{y}\,)$, where A is an equality of two terms, that is, polynomials with natural number coefficients (see [27] and [28]). We say that these formulae are *Diophantine*.

The proof of Theorem 4 depends on the chosen model of computations in the definition of a c.e. set. If Turing machines are used, then the problem reduces to the construction of a Δ_0-definition of the predicate $T(e, x, y)$: "y encodes the computation protocol of the machine with the index e on the input x." This is even simpler than a Gödel arithmetization of the provability predicate in PA, because the rules of operation of the Turing machine are more elementary than the syntax of first-order logic.

Using some technical inventions of Quine, Smullyan [16] developed one of the simplest methods of arithmetization which enables one to avoid the use of both Gödel's β-function and the Chinese remainder theorem. Instead of Turing machines, he considered another useful universal model of computations, namely, the so-called *elementary formal systems* related to Post's canonical systems.

3.4. Semantic version of Gödel's first theorem. As we know, Gödel avoided semantic notions when formulating his theorems. The semantic versions of Gödel's first theorem, which assert the existence of a true unprovable statement under certain conditions, have more natural formulations and simpler proofs. Two things are necessary to pay for this simplicity. First, the syntactic versions of Gödel's theorem are stronger than the semantic ones as a rule. Second, we use a non-elementary notion of truth value in a model. In particular, we cannot even formulate a semantic version of Gödel's theorem in the language of arithmetic.

As we shall see below, the syntactic versions of Gödel's theorem can be obtained from semantic ones by using an additional technical idea.

We recall that a theory is said to be *arithmetical* if an interpretation of the language of Peano arithmetic is fixed in the theory. We may assume that the language of arithmetical theories contains a distinguished sort of natural number variables and all the symbols of the signature of arithmetic for this sort.

Definition 7. Let Γ be an arbitrary set of arithmetical formulae. An arithmetical theory T is said to be Γ-*complete* if for any sentence $A \in \Gamma$ we have

$$\mathbb{N} \vDash A \implies T \vdash A.$$

A theory T is said to be Γ-*sound* if the converse implication holds, that is,

$$T \vdash A \implies \mathbb{N} \vDash A$$

for any sentence $A \in \Gamma$. If Γ is the set of all arithmetical sentences, then the theory T is said to be *semantically complete* and *sound*.

We note that the Γ-completeness of a theory is inherited by its extensions and the Γ-soundness is inherited under passage to a subtheory. Since all axioms of PA are true in the standard model, it is obvious that PA is sound and thus Γ-sound for any Γ.

Lemma 1. *Let T be an arithmetical theory.*

(i) *T is Σ_{n+1}-complete if and only if it is Π_n-complete.*

(ii) *T is Π_{n+1}-sound if and only if it is Σ_n-sound.*

Proof. (i) If $\varphi \in \Pi_n$ and $\mathbb{N} \vDash \exists \vec{x}\, \varphi(\vec{x})$, then $\mathbb{N} \vDash \varphi(\underline{\vec{n}})$ for some tuple $\vec{n}$. If T is Π_n-complete, then $T \vdash \varphi(\underline{\vec{n}})$, and hence $T \vdash \exists \vec{x}\, \varphi(\vec{x})$ by the rules of the predicate calculus. The proof of (ii) is similar.

Theorem 4 enables us to obtain Gödel's first incompleteness theorem in the following standard (semantic) formulation.

Theorem 5. *Let T be a c.e. arithmetical theory.*
(i) *If T is consistent, then T is Π_1-incomplete.*
(ii) *If T is at the same time Σ_1-sound, then there is a Π_1-sentence which is unprovable and irrefutable in T.*

Proof. Suppose that T is Π_1-complete. Then by the preceding lemma, all true Π_1-sentences are provable and all false Π_1-sentences (these correspond to the true Σ_1-sentences) are refutable in T. If T is consistent, then this implies that

$$\mathbb{N} \vDash \varphi \iff T \vdash \varphi$$

for any $\varphi \in \Pi_1$. Since T is c.e., this means that the set of true Π_1-sentences is c.e., which contradicts Corollary 1.

Suppose that T is Σ_1-sound and φ is a true Π_1-sentence which is unprovable in T. Then $\neg\varphi$ is equivalent to a false Σ_1-sentence, and hence $T \nvdash \neg\varphi$. This proves the theorem.

We note that the assumption originally used by Gödel and asserting that T contains the arithmetical axioms is absent in Theorem 5. In particular, the theorem can also be applied to the pure theory of equality in the arithmetical language. However, without additional assumptions we cannot assert that the theory T is undecidable.

Example 1. Consider the model of the language of arithmetic with the support $\mathbb{N}$ in which the symbols 0, S, and $=$ have the ordinary meaning and the symbols $x+y$ and $x \cdot y$ are understood as $\max(x, y)$. Let T be the elementary theory of this model. Of course, T is not sound (in the sense of the standard model), but it is consistent and syntactically complete. Since max can be expressed by using $<$ and the elementary theory $(\mathbb{N}, <)$ is decidable, it follows that T is decidable as well.

This example shows that Theorem 5 is not a consequence of the above abstract syntactic version of Gödel's first theorem.

3.5. Σ_1-completeness and the syntactic version of Gödel's theorem. The passage from the semantic to the syntactic version of Gödel's theorem uses the property of Σ_1-completeness, which can be established for a broad class of arithmetical theories.

Let us first specify the notion of expressibility (Definition 5) in the case of arithmetical theories.

Definition 8. A set $A \subseteq \mathbb{N}$ is said to be *representable*[9] *in* T, or *T-representable*, if there is a formula $\varphi_A(x)$ such that

$$n \in A \iff T \vdash \varphi_A(\underline{n}).$$

We note that T-representability of a set A implies its expressibility in T, because the substitution function $n \mapsto \varphi(\underline{n})$ is computable.

[9]This notion is also referred to as *numerability in* T, and the formula φ_A as a numeration of the set A in T [25].

Lemma 2. *If a theory T is Σ_1-complete and Σ_1-sound, then every c.e. set is T-representable.*

Proof. Let A be a c.e. set and let φ_A be its Σ_1-definition in the standard model in the sense of Definition 4. Then it is clear that φ_A represents A in T. This proves the lemma.

This lemma enables one to apply the above abstract version of Gödel's first theorem if the Σ_1-completeness of the theory T is established.

We also note that for Σ_1-complete theories the semantic condition of Σ_1-soundness can be replaced by a weakened version of Gödel's condition of ω-consistency.

Lemma 3. *A Σ_1-complete arithmetical theory T is Σ_1-sound if and only if for any formula $\varphi(x) \in \Delta_0$ the following conditions do* not *hold simultaneously*:

(i) $T \vdash \exists x\, \varphi(x)$;

(ii) $T \vdash \neg\varphi(\underline{n})$ *for all* $n \in \mathbb{N}$.

The condition in the lemma is usually called 1-*consistency* of the theory T.

It remains to find sufficient conditions for the Σ_1-completeness of an arithmetical theory. To this end, it is sufficient to indicate some Σ_1-complete theory which is as weak as possible. It is customary to mention Robinson's arithmetic Q or even the weaker theory R as such a theory (this tradition goes back to the monograph of Tarski, Mostowski, and Robinson [21]). The theory Q is stronger than R, but it is given by finitely many axioms.

Theorem 6. *The theory Q is Σ_1-complete.*

The idea of proving the Σ_1-completeness is simple, namely, the truth value of any Δ_0-statement φ can be effectively verified. In essence, this verification is a proof of φ in Q if φ is true, or a disproof of φ if φ is false. In fact, for this it is sufficient to show that some simple facts involving specific natural numbers are provable in Q.

Lemma 4. *The following formulae are provable in* Q:

R1. $\underline{m} + \underline{n} = \underline{m+n}$;

R2. $\underline{m} \cdot \underline{n} = \underline{m \cdot n}$;

R3. $\neg(\underline{m} = \underline{n})$ *if* $m \neq n$;

R4. $x \leqslant \underline{n} \;\leftrightarrow\; (x = 0 \vee x = \underline{1} \vee \cdots \vee x = \underline{n})$.

By this lemma one easily proves completeness of Q with respect to Δ_0-sentences, which implies also Σ_1-completeness by Lemma 1.

We refer to the theory in the language of arithmetic extended by the symbol $\leqslant$, axiomatized by all formulae of the form R1–R4, as *Robinson's weak arithmetic* R_0. Thus, Σ_1-completeness also holds for all extensions of R_0. This gives the following corollary.

Corollary 2. *Let T be a Σ_1-sound theory containing* Q (*or even* R_0). *Then all c.e. relations are T-representable.*

Theorem 2 implies the following theorem, which can be regarded as a standard syntactic formulation of Gödel's first incompleteness theorem.

Theorem 7. *Let T be an arithmetical theory such that*

(i) *T contains* Q *(or even* R_0*),*
(ii) *T is c.e.,*
(iii) *T is Σ_1-sound (or 1-consistent).*
Then T is undecidable and incomplete.

We note that the conclusion of this theorem is stronger than the second part of Theorem 5, because we assert in addition that T is undecidable.

One can easily derive the second part of Theorem 5 from Theorem 7. If T is Σ_1-sound, then so is the extension of T by all the axioms of R_0 (the axioms of R_0 are quantifier-free), and Theorem 7 can be applied to this extension.

We also note that Corollary 2 can be established for all consistent theories T [8].

3.6. Gödel–Rosser theorem. The abstract version of the Gödel–Rosser theorem is based on the property that the theory T separates any pair of disjoint c.e. sets. To prove this property, the following lemma is needed.

Lemma 5. *For any $n \in \mathbb{N}$ the formula*

$$\underline{n} \leqslant x \vee x \leqslant \underline{n}$$

is derivable in Q.

We denote by R the extension of R_0 by these formulae for all n. Thus, R is contained in Q. The following lemma asserts that the theory R, and hence also the theory Q, separates every pair of disjoint c.e. sets.

Lemma 6. *Let $A, B \subseteq \mathbb{N}$ be disjoint c.e. sets. Then there is a Σ_1-formula $\varphi(x)$ such that*
(i) $n \in A \Rightarrow \mathsf{R} \vdash \varphi(\underline{n})$,
(ii) $n \in B \Rightarrow \mathsf{R} \vdash \neg\varphi(\underline{n})$
for any $n \in \mathbb{N}$.

Proof. By Theorem 4, there are Δ_0-formulae A_0 and B_0 such that

$$\begin{aligned} n \in A &\iff \mathbb{N} \vDash \exists x\, A_0(\underline{n}, x), \\ n \in B &\iff \mathbb{N} \vDash \exists y\, B_0(\underline{n}, y). \end{aligned}$$

For any formula C and any term t we write

$$\forall x < t\, C(x) \quad \overset{\text{def}}{\iff} \quad \forall x \leqslant t \,\big(x = t \vee C(x)\big).$$

We now set

$$\varphi(z) \quad \overset{\text{def}}{\iff} \quad \exists x \,\big(A_0(z, x) \wedge \forall y < x \neg B_0(z, y)\big).$$

Informally, $\varphi(z)$ asserts that the work of the algorithm receiving the set A on the input z is terminated earlier than the work of the algorithm receiving B is terminated ('Rosser's witness comparisons').

If $n \in A$, then the formula

$$A_0(\underline{n}, \underline{m}) \wedge \forall y < \underline{m} \neg B_0(\underline{n}, y)$$

is true for some m. By the Σ_1-completeness of the arithmetic R, we see that this formula is provable in R, and hence $\mathsf{R} \vdash \varphi(\underline{n})$.

If $n \in B$, then the formula

$$B_0(\underline{n}, \underline{m}) \wedge \forall y \leqslant \underline{m} \neg A_0(\underline{n}, y) \tag{4}$$

is true for some m. Since R is Σ_1-complete, we see that this formula is provable in R. This implies that $\mathsf{R} \vdash \neg\varphi(\underline{n})$. We clarify this assertion by the following reasoning, which can readily be transformed into a formal derivation of a contradiction from the hypothesis that $\varphi(\underline{n})$ holds in R.

Assume $\varphi(\underline{n})$. Then for some x we have

$$A_0(\underline{n}, x) \wedge \forall y < x \neg B_0(\underline{n}, y).$$

If $x \leqslant \underline{m}$, then we have $\neg A_0(\underline{n}, x)$ by (4), which contradicts $A_0(\underline{n}, x)$. If $\underline{m} < x$, then we have $\neg B_0(\underline{n}, \underline{m})$, which contradicts $B_0(\underline{n}, \underline{m})$ by (4). It follows from Lemma 4 that $\forall x\, (x \leqslant \underline{m} \vee \underline{m} < x)$ is derivable in R, which implies the desired contradiction.

This proves the lemma.

The following statement follows immediately from this lemma together with Theorem 3.

Corollary 3. (i) *The system* R *is inseparable.*

(ii) *Every consistent arithmetical theory containing* R *is undecidable.*

This also implies the *Gödel–Rosser theorem.*

Theorem 8. *Let* T *be an arithmetical theory such that*

(i) T *contains* R,

(ii) T *is c.e.*,

(iii) T *is consistent.*

Then T *is inseparable, and thus undecidable and incomplete.*

3.7. Effectiveness of Gödel's and Rosser's theorems. At first glance, the above proofs of Gödel's and Rosser's theorems do not yield explicit examples of independent arithmetical statements, because they are based on arguments by contradiction. However, these proofs can easily be modified so as to produce such examples.

We recall that a standard c.e. undecidable set

$$K = \{x \in \mathbb{N} : \varphi_x(x) \text{ is defined}\}$$

is *creative*, that is, for any c.e. set

$$W_n = \{x \in \mathbb{N} : \varphi_n(x) \text{ is defined}\}$$

one can indicate, effectively with respect to n, a number x such that $x \notin K \cup W_n$ whenever $K \cap W_n = \varnothing$. In fact, for x one can take the number n itself, because $n \in K \iff n \in W_n$.

Let $\psi_K(x)$ be a Σ_1-formula numerating K in T (or expressing K in the standard model of arithmetic). The proofs of Theorems 5 and 7 were based on the fact

that, under the assumption that T is complete, the formula $\neg\psi_K(x)$ numerates the complement of the set K. However, without this (false) assumption we can only assert that the formula $\neg\psi_K(x)$ numerates some c.e. subset K' of the complement of K. Some index of this set, that is, a number n for which $K' = W_n$, can be found explicitly, because we know both the formula $\neg\psi_K(x)$ and the program enumerating the theorems of T. Using the creativity of K, we obtain a number m for which $m \notin K \cup K'$, which implies that neither $\psi_K(\underline{m})$ nor $\neg\psi_K(\underline{m})$ are provable in T. Thus, applying the creativity of K enables us to obtain an effective version of Gödel's first theorem.

To analyze the Gödel–Rosser theorem, we use an effective version of the notion of an inseparable pair of c.e. sets. As we know, pairs of disjoint c.e. sets are identified with computable partial functions $g\colon \mathbb{N} \to \{0,1\}$. A standard example of an inseparable pair of c.e. sets is given by the computable function F in (3), which cannot be extended to an everywhere defined computable function. In fact, this function has the stronger property of *effective inextendibility*, namely, for any computable function φ_n extending F one can, effectively with respect to n, indicate an m such that $\varphi_n(m)$ is undefined.[10] (The corresponding pair of c.e. sets is said to be *effectively inseparable*.)

The proof of the Gödel–Rosser theorem is based on Lemma 5, which is effective in the following sense: for any given pair of disjoint c.e. sets (A, B) this lemma enables us to explicitly indicate a formula ψ separating the pair (A, B) in the theory R. In our case we consider a pair (A, B) defined by the Turing machine M computing the effectively inextendible partial function F. Using the fact that the construction of the Σ_1-definitions of the sets A and B by the machine M is effective, we obtain the corresponding formula ψ_M.

We now associate another computable partial function with the formula ψ_M:

$$f(n) = \begin{cases} 1 & \text{if } T \vdash \psi_M(\underline{n}), \\ 0 & \text{if } T \vdash \neg\psi_M(\underline{n}), \\ \text{undefined} & \text{otherwise.} \end{cases}$$

By the definition of ψ_M we know that the function f extends F. We note that we can effectively find the index n of the function f from the formula ψ_M and the program enumerating the theorems of T. The effective inextendibility of F gives a number m for which the value $f(\underline{m})$ is undefined, that is, neither $\psi_M(\underline{m})$ nor $\neg\psi_M(\underline{m})$ are provable in T.

4. On the limits of applicability of Gödel's first theorem

Both the semantic and the syntactic forms of Gödel's theorem contain several conditions. It is convenient to study them by fixing some conditions and varying others. The most important condition is the enumerability of the theory T, and therefore we consider the role of other conditions under the assumption that T is c.e.

We note that under this assumption the incompleteness of the theory T is derived from the stronger condition of undecidability of T. Thus, the investigation of the

[10] For our function F we have in fact that $m = n$.

question as to whether T is incomplete becomes the investigation of conditions ensuring the undecidability of T.

The questions of decidability and undecidability for logical theories are central in mathematical logic. In particular, great attention was paid to these problems in the works of Mal'tsev and his school. In the present paper we cannot give a complete survey of this topic, but only mention the monograph [29] and the remarkable survey [30], which maintains its relevance to the present day. Here we note only some directions of investigation which are closest to the problem under consideration.

The *semantic formulation* of Gödel's theorem for c.e. theories reduces to the assertion that the set of true arithmetical Π_1-sentences is not enumerable. Possible variations of this theorem go in two directions.

First, by staying in the framework of the language of arithmetic, one can study narrower classes of formulae than Π_1 for which non-enumerability is preserved. The reason for these investigations is to look for undecidable problems of the simplest form and, correspondingly, for the simplest independent sentences. The most interesting results in this direction are connected with the investigations of Hilbert's tenth problem and with the so-called Diophantine forms of Gödel's theorem (for some details, see below).

Second, decidability problems for theories in other languages have been studied very actively. Decidable fragments of the language of arithmetic have been studied especially thoroughly. Beginning with the classical results of Presburger and Skolem, who proved the decidability of the elementary theories of $(\mathbb{N}; =, +)$ and $(\mathbb{N}; =, \cdot)$, respectively, important contributions in this direction were made by the works of J. Robinson, A. Büchi, M. Rabin, A. Woods, Yu. Matiyasevich, A. Semenov, A. Muchnik, J. Richard, P. Cegielski, and others. A modern survey of results concerning this topic is given in [31].

The *syntactic formulation* of Gödel's theorem (in Rosser's stronger form) is an immediate consequence of another statement, namely, of the undecidability of any consistent extension of the system R (Corollary 3). Theories with all their consistent extensions undecidable are said to be *essentially undecidable*.

Tarski [21] noted that the property of incompletability of a c.e. theory T not only follows from the condition of its essential undecidability but is in fact equivalent to this property. Indeed, the effective version of the Lindenbaum lemma shows that every decidable theory has some decidable completion. Thus, the investigation of conditions for incompletability of c.e. theories becomes an investigation of conditions for their essential undecidability. We note that the stronger property of inseparability was proved above for the theory R (Corollary 3).

Tarski found the following sufficient condition for undecidability of theories. This condition uses finitely axiomatized essentially undecidable theories like Q.

Proposition 1. *Let S be a finitely axiomatized and essentially undecidable theory. Then every theory T of the same signature which is compatible with S is undecidable.*

Proof. Suppose that T is decidable. Consider the consistent theory $T' = T + S$. Since S is given by finitely many axioms, we have

$$T' \vdash \varphi \iff T \vdash A_S \to \varphi,$$

where A_S is the conjunction of the universal closures of the axioms of S. Thus, the derivability problem in T' reduces to that in T, and hence T' is also decidable, which contradicts the essential undecidability of S. This proves the proposition.

We note that, in particular, this implies Church's theorem on the undecidability of the predicate calculus (in the signature of arithmetic). As was proved by Tarski, this proposition admits a very useful strengthening in terms of interpretations.

Definition 9. We say that a theory S can be *weakly interpreted* in T if S can be interpreted in some theory U compatible with T and in the language of T.

Proposition 2. *Let S be a finitely axiomatized and essentially undecidable theory. If S can be weakly interpreted in a theory T, then T (as well as every subtheory of T in the same language) is undecidable.*

Proof. Let I be an interpretation of S in a theory U compatible with T. We assume that the parameters of I are defined by some formula $\mathrm{Par}_I(\vec{p})$. We can also assume that S is given by a unique sentence A_S. Then

$$\forall \vec{p}\,\big(\mathrm{Par}_I(\vec{p}) \to A_S^I(\vec{p})\big)$$

is provable in U. Let a theory U_0 be given by this sentence. Then U_0 is compatible with T. However, U_0 is also essentially undecidable: indeed, if $U_0 \subseteq V$ and the theory V is consistent, then the set

$$\big\{\varphi : V \vdash \forall \vec{p}\,\big(\mathrm{Par}_I(\vec{p}) \to \varphi^I(\vec{p})\big)\big\}$$

is deductively closed, consistent, and contains S. It remains to apply Proposition 1. This proves Proposition 2.

Thus, to prove the undecidability of theories by the method of interpretations, it becomes important to obtain examples of weak finitely axiomatized essentially undecidable theories. Tarski, Mostowski, and Robinson used the theory Q for this purpose. One can formally weaken the theory Q without losing essential undecidability by replacing the functions $+$ and $\cdot$ by three-place predicates $A(x,y,z)$ and $M(x,y,z)$. The corresponding system Q^- was recently formulated by Grzegorczyk (see [32]). The axioms of the system are the three axioms of Q involving the successor function and also versions of the other axioms in a relational language which do not assume that the functions of addition and multiplication are everywhere defined:

1. $A(x,y,u) \wedge A(x,y,v) \to u = v,\ M(x,y,u) \wedge M(x,y,v) \to u = v$;
2. $A(x,0,x),\ \exists u\,\big(A(x,y,u) \wedge z = S(u)\big) \to A\big(x,S(y),z\big)$;
3. $M(x,0,0),\ \exists u\,\big(M(x,y,u) \wedge A(u,x,z)\big) \to M\big(x,S(y),z\big)$.

Švejdar [32] proved that the theory Q can be interpreted in Q^-, and therefore Q^- is essentially undecidable. However, Q^- (with the axiom defining $\leqslant$ in terms of addition) is deductively weaker than Q, and it is even Σ_1-incomplete. In particular, $\forall x \leqslant 0\ x = 0$ is provable in Q but not in Q^-.

Weak finitely axiomatized essentially undecidable theories are known for the language of concatenation of words in the binary alphabet, for the weak set theory, and so on (see [33]–[36]). For example, weakening the theory of Tarski and Smielew,

and thus strengthening their result, Vaught showed that this is so for the version of set theory in the signature $\in$ with only the two axioms

$$\forall x\, \exists y\, \neg(y \in x), \qquad \forall x, y\, \exists u\, \forall z\, \big(z \in u \leftrightarrow (z \in x \vee z \in y)\big).$$

Up until now, this theory has apparently been the simplest example (with respect to the formulation) of an essentially undecidable finitely axiomatized theory.

Putnam and Ehrenfeucht ([37], [38]) constructed examples of c.e. essentially undecidable theories S for which Proposition 2 fails to hold. On this background, the result of Cobham (see [33]) asserting that Proposition 2 holds for theories such as R and even R_0, which are not finitely axiomatizable, is of interest. We discuss this result.

First, we note that our formulation of the system R differs somewhat from the traditional one (see [21]). In the traditional formulation the symbol $\leqslant$ is introduced as an abbreviation, and the axiom R4 is weakened to the implication from left to right. However, it can readily be seen that R4 is derivable also in the traditional formulation of R.

Second, the following proposition holds (see [39]).

Proposition 3. (i) R *can be interpreted in* R_0.

(ii) R_0 *is essentially undecidable and inseparable.*

Proof. Let us define an interpretation I of the theory R in R_0 by modifying the order relation in R_0 and leaving unchanged the other symbols of the signature. We write

$$x \leqslant_I y \iff^{\text{def}} \big[\big(0 \leqslant y \wedge \forall u\, (u \leqslant y \wedge u \neq y \to S(u) \leqslant y)\big) \to x \leqslant y\big].$$

It is easy to see that the scheme R4 is provable in R_0 for the relation $\leqslant_I$, as is the additional scheme $x \leqslant_I \underline{n} \vee \underline{n} \leqslant_I x$. This establishes (i).

The interpretation I does not contain parameters. Hence, for any sentence φ we have

$$\mathsf{R} \vdash \varphi \implies \mathsf{R}_0 \vdash \varphi^I.$$

Since I preserves the numerals, this immediately implies that R_0, as well as R, separates pairs of c.e. sets, and thus is an inseparable theory. This completes the proof of the proposition.

Corollary 4. *The Gödel–Rosser theorem holds also for all consistent extensions of* R_0.

The theory R_0 can be further weakened. Using J. Robinson's idea showing that addition is expressible in the model $(\mathbb{N}; =, S, \cdot)$, Jones and Shepherdson [39] established that R_0 can be interpreted in the theory R_1 obtained from R_0 by getting rid of both the addition operation and the scheme R1. As above, this implies both the essential undecidability and the inseparability of R_1.

On the other hand, Cobham proved his theorem for a weaker relational version R_0^- of the theory R_0.

The language of R_0^- contains the predicate symbols C_0, Sc, A, M, and $\leqslant$ of arity 1, 2, 3, 3, and 2, respectively, and does not contain equality. The predicate $C_0(x)$ distinguishes the constant 0, and Sc, A, and M define the graphs of the

successor, addition, and multiplication functions. Instead of numerals, we use the formulae $C_n(x)$ inductively defined by $C_{n+1}(x) \overset{\text{def}}{\iff} \exists y \big(C_n(y) \wedge Sc(y,x)\big)$.

The theory R_0^- contains the following axioms:[11]

(i) $\exists x\, C_0(x)$, $\neg\big(C_m(x) \wedge C_n(x)\big)$ for $m \neq n$;
(ii) $C_m(x) \wedge C_n(x) \to \big(A(x,y,z) \leftrightarrow C_{m+n}(z)\big)$;
(iii) $C_m(x) \wedge C_n(x) \to \big(M(x,y,z) \leftrightarrow C_{mn}(z)\big)$;
(iv) $C_n(y) \to \big(x \leqslant y \leftrightarrow C_0(x) \vee C_1(x) \vee \cdots \vee C_n(x)\big)$.

We note that R_0^- does not contain equality axioms, and therefore it is significantly weaker than R_0 and more convenient in applications based on the following theorem of Cobham.

Theorem 9. *If R_0^- can be weakly interpreted in T, then T (and any subtheory of it) is undecidable.*

The essential undecidability of R_0^- follows immediately from the given theorem. However, it can also be established directly in a rather simple way.

Since the language of R_0^- does not contain numerals, the notion of decidability of a k-place relation A in a theory T in the signature of R_0^- can be modified as follows: there is a formula $\varphi_A(x_1,\dots,x_k)$ such that for all $n_1,\dots,n_k$

$$A(n_1,\dots,n_k) \implies T \vdash C_{n_1}(x_1) \wedge \cdots \wedge C_{n_k}(x_k) \to \varphi_A, \tag{5}$$

$$\neg A(n_1,\dots,n_k) \implies T \vdash C_{n_1}(x_1) \wedge \cdots \wedge C_{n_k}(x_k) \to \neg\varphi_A. \tag{6}$$

Using this modification, one can easily prove decidability in R_0^- of any Δ_0-predicates and also separability of pairs of c.e. sets.

Corollary 5. *The theory R_0^- is inseparable and essentially undecidable.*

Cobham's theorem immediately yields one of the strongest forms of the Gödel–Rosser theorem.

Theorem 10. *If R_0^- can be weakly interpreted in a c.e. theory T, then T is undecidable and incomplete.*

Vaught [33] suggested a rather simple derivation of Cobham's theorem from Trachtenbrot's theorem on the inseparability of the set of identically true formulae of the predicate logic and of the set of formulae refutable on finite models. (As the signature, it is sufficient here to take a single binary predicate symbol.) Thus, Trachtenbrot's theorem gives another proof of the Gödel–Rosser theorem.

Visser [40] gave an interesting characterization of theories mutually interpretable with R (or R_0) in terms of satisfiability on finite models.

A theory T is said to be *finitely satisfiable* if T has a finite model. A theory T is said to be *locally finitely satisfiable* if any finite subtheory of T is finitely satisfiable. Obviously, the theory R is locally finitely satisfiable but not finitely satisfiable. Visser showed that R has the following maximality property.

[11] There is a misprint in the definition of R_0^- in [33] and [30], namely, the symbol $\to$ in the axioms (ii) and (iii) must be replaced by $\leftrightarrow$. Such a theory cannot be essentially undecidable, because it can be interpreted in the decidable theory $\mathrm{Th}(\mathbb{N};\leqslant)$ upon translation of A and M by identically false formulae.

Theorem 11. *Every c.e. locally finitely satisfiable theory T is interpretable in* R (*and the interpretation is one-dimensional and parameter-free*).

This theorem shows that, although there is ambiguity in the choice of the signature, the axioms, and some other details in the formulation of R, the theory R, modulo interpretability, nevertheless occupies a privileged position and is distinguished by a certain natural general property.

We also note that, along with arithmetical theories of type R, essentially undecidable locally finitely satisfiable theories in other languages have also been considered. One of the most beautiful examples is the theory S introduced by Vaught [41], defined in the signature of set theory by the single scheme of axioms

$$\forall x_1, \dots, x_n \, \exists y \, \forall z \left(z \in y \leftrightarrow \bigvee_{i=1}^{n} z = x_i \right) \quad \text{for all } n \geqslant 1.$$

Vaught showed that the theory S interprets R_0^-, and thus is inseparable and essentially undecidable.

5. On the proofs of Gödel's first theorem

Many proofs of Gödel's first theorem in the literature follow the general scheme presented above. However, they can differ significantly from one another in technical details. Reproducing all the details leaves (at least) the choice of the following parameters (which we shall discuss for the simpler semantic version of Gödel's theorem) to the taste of the authors:

1. the choice of a specific basic formal system T;
2. the choice of a universal computation model on some family U of objects of T, where by such a model I mean any mathematically rigorous definition of the notion of a c.e. set of elements of U (or of a computable function from U to U);
3. the choice of a Gödel numbering, that is, an encoding of the syntax of the theory T by objects in U;
4. a proof of the enumerability of the system T (in the sense of the chosen computation model and the Gödel numbering);
5. a presentation of an example of an expressible non-c.e. set (together with the proof of its expressibility and non-enumerability).

We note immediately that many authors of simplified proofs of Gödel's theorem neglect some of these points due to their intuitive clearness, using, as a rule, an informal concept of algorithm and some of the forms of the Church–Turing thesis. For example, the enumerability of the arithmetic PA is intuitively clear. At the same time, an 'honest' proof of this statement needs programming in the framework of the chosen computation model, that is, significant technical work in general. Choosing the apparatus of primitive recursive functions, Gödel managed quite effectively with the problem. Of course, the comparison may be valid only for more or less *complete* proofs, although from the pedagogical point of view it is quite justified and, moreover, quite desirable to neglect intuitively clear statements. In our opinion there are still no complete proofs of Gödel's theorem that are essentially simpler than his own proof.

Let us consider some of the above points step by step.

Choice of a theory. As a rule (but not always), authors choose some arithmetical theory for the system T. Diverse versions are possible for pedagogical or ideological reasons, depending on the choice of the signature. For example, it is convenient to develop the encoding in a language having a symbol for the function of raising to a power.

The primitive recursive arithmetic PRA whose language includes terms for all primitive recursive functions is often taken as T. In this system one can work directly with primitive recursive functions and avoid the inconvenience of embedding them in the language of PA [3]. The language of PRA is infinite, although it is primitive recursive. Moreover, PRA is too strong a theory, far beyond the theories of bounded arithmetic, and this complicates the generalization of Gödel's theorem to this important class of formal systems.

Theories of binary labelled trees [42], theories of words in a binary alphabet ([43], [44]), theories of hereditarily finite sets [45], and some other theories have also been considered as an alternative to arithmetical theories. The description of the syntax of formal theories or computation models in these systems is more natural, because it enables one to avoid a Gödel numbering in a sense, namely, the formulae themselves are identified with objects of the theory (words, labelled trees, or finite sets).

Variations in the choice of the proof systems of T are also possible, namely, one can consider natural deduction, Gentzen's sequent calculus, and many other formats of proofs. As was already noted above, these variations are not very fundamental for Gödel's first theorem, and, as a rule, the most customary and simply formulated Hilbert format is used.

Choice of a computation model. To prove Gödel's theorem, various authors have considered the following computation models:

1) c.e. sets as projections of primitive recursive relations (Gödel);
2) the Herbrand–Gödel computable functions and partial recursive functions (Kleene);
3) elementary formal systems (Smullyan);
4) Turing machines;
5) Σ-definable relations (Ershov) and others.

The choice of each of these models has its own merits and drawbacks.

In my opinion, Turing machines have the advantage that they are the most standard and natural computation model. Among all known theoretical models, the Church–Turing thesis is the most convincing for Turing machines. For this reason, these machines are ideal for 'abbreviated' proofs of Gödel's theorem.

Since it is somewhat more complicated to program Turing machines than to use the higher-level language of partial recursive functions, an explicit construction of a universal c.e. set, as well as a proof of enumerability of arithmetics, is not as simple as using the language of partial recursive functions.

The choice of Σ-definability as an independent computation model (Ershov [45]) is of interest. Here, in essence, Theorem 4 becomes the definition of a c.e. set and requires no proof. As a requital for this choice, it becomes necessary to construct a universal c.e. set in the framework of the model. It is easy to see that such a set is given by the formula which is customarily referred to in logic as the Σ-definition

of truth for Σ-formulae (see [26] and [45]). In its technical difficulty the problem of constructing this formula is similar to the problem of an arithmetical definition for the provability predicate. Nevertheless, conceptually, the advantage of the choice of Σ-definability as the computation model is that all proofs are carried out in the framework of the same language, for set theory or formal arithmetic, and no use of any external mechanisms of computability is required.

Among the known 'honest' proofs of Gödel's theorem, Smullyan's proof ([16], [9]) is one of the exemplary ones. It is thorough with respect to both the design and the details. In particular, his choice of elementary formal systems as a model of computability is beautifully adapted to the requirements of formalization of logical languages.

Choice of a non-c.e. set. The choice of a specific non-c.e. set used to construct an independent statement influences the external aspect of the proof most radically (without changing the essence, of course). Sometimes this set is associated with a formal analogue of some semantic paradox. Gödel's own proof (see below) is based on the construction of a formula asserting its own unprovability, which is analogous to the liar's paradox. (We note that if a theory T is semantically complete, then the statement asserting that a formula φ is unprovable in T is equivalent to the falseness of φ.) However, Gödel noted that one can adapt almost any of the known semantic paradoxes to prove his theorems. Many later authors successfully confirmed this idea of the classical author.

Kleene [12] was apparently the first to give the abstract computational core of Gödel's theorem in the spirit of the approach discussed here, using the notion of Herbrand–Gödel general recursive function. For his first recursion-theoretic proof of Gödel's theorem, Kleene used the non-enumerability of the set of all indices of general recursive functions (that is, everywhere defined computable functions). He later noted [13] that this proof is very close to Richard's paradox (and to Cantor's diagonal construction).

Other known proofs are connected with Berry's paradox: does there exist a *least natural number that cannot be defined by fewer than seventeen words of the English language*? Two proofs using a similar idea were suggested by Chaitin [46] and Boolos [47]. Boolos uses the ordinary notion of definability in arithmetic as an explication of the notion 'to define'. We dwell on Chaitin's proof in some detail because it became widely known, also in the popular scientific literature.

His proof uses the notion of the *Kolmogorov complexity* $K(x)$ of a number x, that is, for the 'definitions' of x one considers all possible programs p such that on the empty input the program p produces x. In familiar terms, we can define one of the versions of the function K for the standard (optimal) numeration of Turing machines as follows:

$$K(x) = \min\{n \in \mathbb{N} : \varphi_n(0) = x\}.$$

Since the relation $K(x) > y$ is non-c.e. and expressible in arithmetic, we can conclude that there are unprovable true statements of the form $K(\underline{n}) > \underline{c}$ for some constants n and c. Chaitin asserts a bit more, namely, he presents an entire series of unprovable statements of this form.

Theorem 12. *Let T be a c.e. and sound arithmetical theory. Then*

$$\exists c\, \forall m\, T \nvdash K(\underline{m}) > \underline{c}.$$

The computational basis of this theorem lies in the following well-known property of Kolmogorov complexity: for a given c there is no effective way to find a number m for which $K(m) > c$. (At the same time, it is obvious that for any c such a number m exists, because the number of programs whose indices do not exceed c is finite.)

Lemma 7. *There is no everywhere defined computable function f which satisfies $\forall c\, K(f(c)) > c$.*

Chaitin's theorem follows from this lemma. Indeed, if

$$\forall c\, \exists m\, T \vdash K(\underline{m}) > \underline{c},$$

then since T is c.e., we obtain a computable everywhere defined function f for which $\forall c\, T \vdash K(\underline{f(c)}) > \underline{c}$, and since T is sound, this implies that $\forall c\, K(f(c)) > c$, which contradicts the lemma.

To prove the lemma, we use Kleene's recursion theorem. Arguing by contradiction, we assume that there is an f satisfying the assumptions of the lemma. Consider a computable function g which for an input c first computes $y = f(c)$ and then produces some index n such that $\varphi_n(0) = y$. Using the recursion theorem, we choose an index e such that the function φ_e coincides with $\varphi_{g(e)}$. By the construction of g we have $K(f(e)) \leqslant g(e)$. However, since the output of the program e is just like that of $g(e)$, we also have $K(f(e)) \leqslant e$, which contradicts the condition of the lemma.

The short proof of Chaitin's theorem using the recursion theorem is a modified version of the proof in [48] (see also [49]). Chaitin himself gave a more lengthy but direct argument (see [46], [50], and also [51]). We note that the constant c depends on the chosen numeration of computable functions, and for some choice of this numeration one can achieve the condition $c = 0$ for all theories T [48].

As we see, Chaitin's theorem is an immediate manifestation of the non-computability of some simple properties connected with the Kolmogorov complexity function $K(x)$. Such facts about $K(x)$ were certainly well known, as was the general principle which enables one to convert them into incompleteness results. From this point of view, Chaitin's result is not more interesting than other proofs of the incompleteness theorem that follow the above general scheme.

6. Gödel's proof

From the point of view of computability theory, the analysis of Gödel's original proof and of its later versions (say, of Theorem 7; see, for example, [52] and [15]) is perhaps of most interest. This proof is based on the technical notion of *representability of a function* in a theory T and in the construction of an arithmetical formula asserting its own unprovability. The plan of Gödel's proof can be described as follows:

1) the proof of the fact that the proof predicate $\mathrm{Prf}_T(x, y)$ for the theory T is primitive recursive;

2) the proof of the fact that every primitive recursive function is representable in T, which implies the decidability in T of the predicate $\mathrm{Prf}_T(x, y)$;
3) the construction of a formula ψ such that

$$T \vdash \psi \leftrightarrow \neg \mathrm{Pr}_T\big(\underline{\ulcorner \psi \urcorner}\big),$$

where $\mathrm{Pr}_T(x)$ stands for the provability formula $\exists y\, \mathrm{Prf}_T(x, y)$, and where we use the representability in T of the substitution function.

The proof is completed with the following argument, which shows that if T is ω-consistent, then the formula ψ is unprovable and irrefutable in T.

If $T \vdash \psi$, then $\mathbb{N} \vDash \mathrm{Prf}_T\big(\underline{\ulcorner \psi \urcorner}, n\big)$ for some n, which yields $T \vdash \mathrm{Prf}_T\big(\underline{\ulcorner \psi \urcorner}, \underline{n}\big)$, because Prf_T is decidable in T, and thus $T \vdash \mathrm{Pr}_T\big(\underline{\ulcorner \psi \urcorner}\big)$. By the definition of ψ, this implies that $T \vdash \neg\psi$, in other words, T is inconsistent.

If $T \vdash \neg\psi$, then $T \vdash \exists y\, \mathrm{Prf}_T\big(\underline{\ulcorner \psi \urcorner}, y\big)$ by the definition of ψ. On the other hand, since T is consistent, we have $T \nvdash \psi$. Thus, for any specific n we have $\mathbb{N} \vDash \neg\mathrm{Prf}_T\big(\underline{\ulcorner \psi \urcorner}, n\big)$. Since the predicate Prf_T is decidable in T, we see that $T \vdash \neg\mathrm{Prf}_T\big(\underline{\ulcorner \psi \urcorner}, \underline{n}\big)$ for any n, that is, T is ω-inconsistent.

We note that the step 1) of the proof is one of the historically first experiences in programming (in this case, in the language of primitive recursive functions). Let us consider the steps 2) and 3) in greater detail.

A function $f(\vec{x}\,)$ is said to be *representable in* T if the following is satisfied for some formula $\phi_f(\vec{x}, y)$: for any $\vec{n}$ if $m = f(\vec{n}\,)$, then

$$T \vdash \forall y\, \big(\varphi_f(\underline{\vec{n}}, y) \;\leftrightarrow\; y = \underline{m}\big).$$

The function f is said to be Σ_1-*representable* if φ_f can be chosen in the class of Σ_1-formulae. The following lemma holds.

Lemma 8. *Every computable function f is Σ_1-representable in* R.

As a rule, this lemma is proved directly, by induction with respect to the construction of a partial recursive scheme defining f. It can also easily be derived from Corollary 2, using a modification of the Σ_1-definition of the graph of the function f. If a formula $\exists z\, \varphi_0(\vec{x}, y, z)$ represents the graph of f, where $\varphi_0 \in \Delta_0$, then it suffices to take φ_f to be a formula expressing the fact that y is the first element of a minimal pair $\langle y, z\rangle$ for which $\varphi_0(\vec{x}, y, z)$ holds.

We note that the representability of the characteristic function of a set A in T implies the decidability of A in T, and therefore Lemma 8 implies the decidability in R of any algorithmically decidable set. Considering partial $\{0, 1\}$-valued functions, we can similarly derive from Lemma 8 the representability in R of any c.e. set and the separability of pairs of c.e. sets.

Fixed-point lemma. The central point of Gödel's proof was the construction of a sentence ψ asserting its own unprovability. The possibility of constructing arithmetical sentences referring to themselves does not depend on the specific features of the provability formula and in fact holds for any property expressible by a formula of the language under consideration. In modern presentations the corresponding statement is distinguished as a separate *fixed-point lemma*.

Let T be an arithmetical theory and let $\varphi(x)$ be a formula of the language of T with a numerical variable x, where $\varphi(x)$ may possibly contain other free variables.

For any formula ψ of the language of T denote by $\varphi[\psi]$ the formula $\varphi(\underline{m})$, where $m = \ulcorner\psi\urcorner$ stands for the Gödel number of ψ. As always, we assume that the Gödel numbering of the formulae is computable.

Lemma 9. *Let T be an arithmetical theory containing* R. *Then for any formula $\varphi(x)$ of the language of T there is a formula ψ with the same free variables besides x as those of φ such that*

$$T \vdash \psi \leftrightarrow \varphi[\psi].$$

Proof. Consider the function assigning to any formula $\theta(x)$ the formula $\theta[\theta]$. This function is computable, because the Gödel numbering of the formulae is computable. Hence, the function

$$\ulcorner\theta\urcorner \mapsto \ulcorner\theta[\theta]\urcorner$$

is also computable. Let $\mathrm{Diag}(x, y)$ denote a Σ_1-formula representing this function in R. Consider the formula

$$\varphi'(x) = \exists y \big(\varphi(y) \wedge \mathrm{Diag}(x, y)\big).$$

Let $\psi = \varphi'[\varphi']$. By the definition of Diag we have

$$\mathsf{Q} \vdash \forall y \big(\mathrm{Diag}(\underline{\ulcorner\varphi'\urcorner}, y) \leftrightarrow y = \underline{\ulcorner\psi\urcorner}\big).$$

By the rules of the predicate logic with equality, this implies that the formula $\varphi'[\varphi']$ is equivalent to $\exists y \big(\varphi(y) \wedge y = \underline{\ulcorner\psi\urcorner}\big)$, that is, $\varphi[\psi]$, as was to be proved. This completes the proof of the lemma.

Analysis of Gödel's proof. Let us consider Gödel's proof from the computational point of view. The provability in a Σ_1-sound c.e. theory T containing R provides us with another universal model of computability (Corollary 2). Therefore, one may *define* c.e. sets as the sets that are numerable in T. According to this point of view we may treat a formula $\varphi_A(x)$ numerating A in T as a *program* accepting the set A, namely,

$$n \in A \iff T \vdash \varphi_A(\underline{n}).$$

The Gödel number of any formula φ_A of this kind is regarded as an index of the set A. In this case the computation of the program φ_A on the input n is the search for a derivation of the formula $\varphi_A(\underline{n})$ in T.

For this computation model one can reformulate all general results known in computability theory for any standard model, for example, for Turing machines. We recall that a creative set K was defined as $\{m \in \mathbb{N} \mid m \in W_m\}$. Here the set

$$\{n \in \mathbb{N} \mid T \vdash \varphi_A(\underline{n})\}$$

is used as W_m, where $\ulcorner\varphi_A\urcorner = m$. This is equivalent to

$$\big\{n \in \mathbb{N} \mid \mathbb{N} \vDash \exists z \big(\mathrm{Pr}_T(z) \wedge \mathrm{Sub}(m, n, z)\big)\big\},$$

where $\mathrm{Sub}(x, y, z)$ represents the function $\ulcorner\varphi\urcorner,\ n \mapsto \ulcorner\varphi(\underline{n})\urcorner$. Correspondingly, the analogue of the set K is numerated by the formula

$$\exists z \big(\mathrm{Pr}_T(z) \wedge \mathrm{Sub}(x, x, z)\big). \tag{7}$$

Repeating the effective version of the proof of Theorem 7, we consider a c.e. set numerated by the negation of the formula (7) and compute its index m (the Gödel number). By the definition of K, the sentence

$$\neg\exists z\,\big(\mathrm{Pr}_T(z) \wedge \mathrm{Sub}(\underline{m},\underline{m},z)\big)$$

is independent. The attentive reader has certainly already noticed that the formula $\mathrm{Sub}(x,x,y)$ is simply $\mathrm{Diag}(x,y)$, and we have just literally written out a solution of the fixed-point equation

$$T \vdash \psi \leftrightarrow \neg\mathrm{Pr}_T(\ulcorner\psi\urcorner)$$

given in the proof of Lemma 9. Thus, Gödel's proof of the incompleteness theorem is equivalent in essence to an effective version of a recursion-theoretic proof. The only difference is that Gödel chose a specific model of computability.

The above analysis also clarifies the role of the notion of representable function. We could manage without this notion also for the proof of Theorem 7 by using Gödel's approach and by weakening the formulation of the fixed-point lemma. If T is Σ_1-sound, then for our purposes it is sufficient that the equivalence $\psi \leftrightarrow \neg\mathrm{Pr}_T(\ulcorner\psi\urcorner)$ is true in the standard model of arithmetic, and we can obtain this fact already from the simple numerability of the relation Sub by some Σ_1-formula. It is quite another point that a stronger formulation enables us to assert the unprovability of the formula ψ already under the assumption of simple consistency of T, and this is essential, for example, when proving Gödel's second theorem.

We note that the original proof of the Gödel–Rosser theorem [18] is also based on an application of the fixed-point lemma. Rosser considered the following *Rosser's proof predicate* for T:

$$\mathrm{Prf}^R_T(x,y) \quad\stackrel{\mathrm{def}}{\Longleftrightarrow}\quad \mathrm{Prf}_T(x,y) \wedge \forall z<y\,\forall u<z\,\big(\mathrm{Neg}(x,u) \to \neg\mathrm{Prf}_T(u,z)\big),$$

where $\mathrm{Neg}(x,u)$ represents the function of negating the formula x: $\ulcorner\varphi\urcorner \mapsto \ulcorner\neg\varphi\urcorner$. We note that if the theory T is consistent, then $\mathrm{Prf}^R_T(x,y)$ defines in the standard model the same predicate that is defined by Gödel's formula $\mathrm{Prf}_T(x,y)$. Rosser's provability formula is defined by $\mathrm{Pr}^R_T(x) \stackrel{\mathrm{def}}{\Longleftrightarrow} \exists y\,\mathrm{Prf}^R_T(x,y)$.

Following Rosser, one can easily prove that if

$$T \vdash \theta \leftrightarrow \neg\mathrm{Pr}^R_T[\theta], \tag{8}$$

then the formula θ is neither provable nor refutable in T already under the assumption of simple consistency of T.

This proof can also be reduced to a recursive-theoretic proof for an appropriate model of computability. In the present case we work with the notion of computable $\{0,1\}$-valued partial recursive function and regard a formula $\varphi(x)$ as a program for the computation of the function

$$f(n) = \begin{cases} 0 & \text{if } T \vdash \varphi(\underline{n}),\\ 1 & \text{if } T \vdash \neg\varphi(\underline{n}),\\ \text{undefined} & \text{otherwise.}\end{cases}$$

An analysis similar to that above shows that the independent statement obtained from the effective version of Rosser's theorem corresponds literally to the solution θ of the fixed-point equation (8).

7. Incompleteness theorem and algorithmic problems

Using the above general scheme for proving Gödel's theorem, one can convert every known undecidable algorithmic problem in mathematics into an incompleteness theorem. As a typical illustration, we consider several examples in arithmetic and group theory.

Hilbert's tenth problem: decide for a given Diophantine equation whether it has at least one integer solution. This example is related to arithmetic and is therefore especially close to the traditional version of Gödel's theorem. It is of interest because it gives the simplest form (in the logical sense) of an independent statement (for the standard language of arithmetic).

By the Matiyasevich theorem, the set K is expressible by some Diophantine formula of the form $\exists\vec{y}\, p(x,\vec{y}\,) = q(x,\vec{y}\,)$, where p and q are polynomials with natural number coefficients. The negation of this formula defines a non-c.e. set, and hence for a given c.e. consistent arithmetical theory T we obtain a constant c for which the statement $\forall\vec{x}\, p(\underline{c},\vec{y}\,) \neq q(\underline{c},\vec{y}\,)$ is true but not provable in T. The authors of [53], using earlier results of Jones, gave a quite transparent explicit formula (easily convertible into a Diophantine formula) for which the above statement is true. The constant c can also be given explicitly if a Turing machine enumerating T is specified and an arithmetization of Turing machines is fixed.

Word problem in groups. Let a presentation of some group G by a finite system of generators and defining relations be given. For a given word w in the group alphabet

$$\{a_1, a_1^{-1}, \ldots, a_n, a_n^{-1}\}$$

formed by the generators and their inverses it is required to determine whether or not the equality $w = 1$ holds in G. A classical result of Novikov [54] is the construction of a finitely presented group G for which this algorithmic problem is undecidable.

Using the standard Gödel numbering of words in the alphabet of G, we can define in the language of arithmetic the c.e. undecidable predicate $w =_G 1$ expressing the equality in G of a word w to the identity element. According to the general scheme, for a given (c.e. and consistent) theory T we obtain a specific word $w_T \neq_G 1$ for which this fact is not provable in T.

Unrecognizability of invariant properties of a group from a finite presentation of it. By the well-known Adian–Rabin theorem (see [55] and [56]), almost all non-trivial invariant properties of groups are unrecognizable from finite presentations of them. For example, this is the case for the property that a group is isomorphic to the trivial group.

Consider a natural Gödel numbering of all finite presentations of groups by generators and defining relations. Generalizing the previous example, we can express in the language of arithmetic the c.e. two-place predicate $w =_G 1$, where G is a group with a finite presentation and w is a word in the alphabet of G. We note that

the property that G defines the identity group is c.e., because it is equivalent to the statement $\bigwedge_{i=1}^{n}(a_i =_G 1)$, where $a_1, \ldots, a_n$ are the generators of G. Thus, the *non-identity* property of a group is non-c.e. and is expressible by an arithmetical Π_1-formula. The general scheme for proving Gödel's theorem for any c.e. theory T provides us with an example of a finitely presented non-identity group G_T whose non-identity is not provable in T if T is consistent.

Examples of finitely presented infinite groups whose infiniteness is not provable in T, torsion-free groups for which this fact is not provable in T, and so on, can be constructed in a similar way. The list of examples of this kind can be extended. Like Chaitin's theorem, these examples do not add much to our understanding of provability in formal systems. On the other hand, the examples show the presence of unprovable facts touching on very diverse areas of mathematics, including areas that are very far from arithmetic.

A typical unprovable statement A is of the form "a given object C has the property P," where the property P can be quite natural mathematically. However, specific objects C for which A is unprovable are rather large as a rule, because practically all known examples of non-c.e. sets are based on encoding of programs (for some universal computation model). Correspondingly, the examples of unprovable statements thus obtained depend on rather large constants. In turn, these constants depend on a specific Gödel numbering of the programs, for example, of Turing machines, and therefore are by no means 'canonical'. This somewhat reduces the value of these examples from the point of view of ordinary mathematics.

The question of whether these constants can still be used for some meaningful classification of formal systems, for example, after fixing some specific 'natural' Gödel numbering, is of interest. Along with Chaitin, other authors also have speculated on this topic (see, for example, [53]). The author of the present survey knows no working examples of such classifications, but in principle this question can be assumed to be open for now.[12]

8. Mathematically natural examples of unprovable statements

The situation is rather different with natural examples of unprovable statements found in diverse areas of mathematics during the period after Gödel discovered his theorems.

The most famous example of a statement of this kind was Cantor's continuum hypothesis. The fact that this hypothesis does not contradict the axioms of the set theory ZFC was proved by Gödel, and the unprovability of this hypothesis in ZFC was proved by Cohen (both results were obtained under the assumption that ZFC is consistent). Independence from the axioms of ZFC was later established for many other statements in set theory, general topology, and the theory of functions. All these results were obtained on the basis of other, deeper approaches connected with specific features of set theory. In particular, natural independent

[12]Examples in which one can make all constants vanish for some (artificial) choice of a Gödel numbering in Chaitin's theorem show that attempts at classifications of this kind cannot be quite naive. However, this does not exclude the possibility that there are more complicated solutions of the problem.

statements in set theory strongly involve infinite sets, which fundamentally distinguishes these statements from Gödel's independent formulae having a finitary[13] nature (and expressible in the language of arithmetic).

Natural mathematical statements of a finitary (as a rule, combinatorial) nature that are independent of standard logical theories were discovered much later. Here the theories were taken to be Peano arithmetic PA, the second-order predicative arithmetic ATR_0, or even the set theory ZFC together with some axioms of large cardinals. Each example of a finitary independent statement is based on a deep analysis of a specific formal system and does not generalize to arbitrary c.e. theories. In this sense, we are concerned here with independence results of another, *non-Gödelian type*.

It should be noted that these results imply the incompleteness of a number of important specific theories, for example, of PA or even ZFC. However, it is impossible to derive the fact of their *fundamental incompletability*, that is, Gödel's first theorem, at the same level of generality as given by Gödel himself. Nevertheless, this loss of generality is more than compensated by the naturalness of the statements whose independence is established. Below we describe only some of the most well-known examples of this kind, but a detailed survey of the topic is far beyond the framework of this paper (for a readable introduction see, for example, the paper [57]).

Paris–Harrington principle. One of the first and most striking examples of a mathematically natural finitary independent statement was the so-called *Paris–Harrington principle* generalizing the finite Ramsey theorem and found in [58].

We denote by $[X]^k$ the set of all k-element subsets of a finite set X and by $|X|$ the cardinality of X. For X we shall take an initial segment $\{0, 1, \dots, n-1\}$ of the natural numbers, which we denote by $\mathbf{n}$. We consider colourings of the set $[X]^k$ with c colours, that is, functions $f\colon [X]^k \to \mathbf{c}$. A subset $Y \subseteq X$ is said to be *f-homogeneous* if the restriction of the function f to the subset $[Y]^k \subseteq [X]^k$ is constant.

The classical Ramsey theorem asserts that for any c, k, and m there is an n such that for any colouring $f\colon [\mathbf{n}]^k \to \mathbf{c}$ there exists a one-colour subset $Y \subseteq \mathbf{n}$ for which $|Y| \geqslant m$. As is well known, this theorem can be formalized and proved in the Peano arithmetic PA or even in the weaker primitive recursive arithmetic.

The Paris–Harrington principle differs from this theorem only by an additional condition on the one-colour set $Y \subseteq \mathbf{n}$ in the conclusion of the theorem, namely, together with the condition $|Y| \geqslant m$, the condition $|Y| > \min(Y)$ must also hold. Like the classical theorem, this principle can easily be derived from the infinite Ramsey theorem by using compactness considerations. It is all the more surprising that, as was shown by Paris and Harrington, their principle is not provable in Peano arithmetic.

Besides the Paris–Harrington principle, several typical non-derivable combinatorial statements are known for Peano arithmetic: 'the Goodstein sequence', 'the Hercules–Hydra game' [59], 'the Kanamori–McAloon principle' [60], 'the Worm

[13] That is, formulated without using infinite objects. Here we use the word *finitary* in a somewhat weaker sense than it is customary in the framework of a discussion of Hilbert's programme.

principle' ([61], [62]), and others. From the logical point of view, all these principles are equivalent to the statement of 1-consistency of the arithmetic PA, and therefore are provable in stronger theories like ATR_0 and ZFC.

Finitary version of Kruskal's theorem. The most well-known example of a finitary combinatorial statement independent of the theory ATR_0 is connected with Kruskal's theorem on trees. This statement was found by Harvey Friedman and published in [63].

Let us consider finite trees $(T, \leqslant, \inf)$ as partially ordered sets with the operation inf assigning to elements $x, y \in T$ their greatest lower bound $\inf(x, y)$. By a *homeomorphic embedding* of a tree T_1 in a tree T_2 we mean an injective map $f\colon T_1 \to T_2$ preserving the operation inf (and thus the order relation): $f\big(\inf(x, y)\big) = \inf\big(f(x), f(y)\big)$.

Kruskal's theorem, in its simplest form, asserts that for any infinite sequence of finite trees $T_0, T_1, \ldots$ there are indices $i < j$ such that T_i can be homeomorphically embedded in T_j. We note that this statement is not finitary, because it appeals to arbitrary infinite sequences (of finite trees).

Consider the following finitary corollary to Kruskal's theorem. For any m there is an n such that if $T_0, T_1, \ldots, T_n$ is a sequence of finite trees in which every tree T_k has at most $m + k$ vertices, then T_i can be homeomorphically embedded in T_j for some $i < j \leqslant n$. This statement follows easily from Kruskal's theorem by compactness considerations. On the other hand, as was proved by Friedman, the statement implies the 1-consistency of ATR_0, and therefore is not provable in ATR_0 by Gödel's second theorem. In fact, the finitary form of Kruskal's theorem goes rather far beyond the framework of ATR_0. In [64] and [65] a precise characterization of the infinite and finite Kruskal theorems is obtained in terms of proof-theoretic ordinals and rapidly growing functions. (A sharp bound is given by the so-called *small Veblen ordinal* and by the function of the extended Grzegorczyk hierarchy with the corresponding index.)

Friedman also found stronger finitary combinatorial statements like Kruskal's theorem that are connected with a specific notion of embedding (so-called *gap embedding*) for labelled finite trees [63]. Other interesting combinatorial principles of similar strength are also known, for example, the so-called 'Buchholz's Hydra game' generalizing the game 'Hercules–Hydra' to labelled trees [66].

Finitary statements independent of ZFC. In subsequent years Friedman obtained a series of new finitary statements that are independent of increasingly stronger theories, including the ZFC theory together with some axioms of large cardinals. (These statements are provable under the assumption that some other, stronger axioms of large cardinals are satisfied.) Some of these results are published in [67], but most of them are accessible only in the form of preliminary communications (see [68]). It is striking that, in contrast to the traditional independence results in set theory, these statements are completely finitary. However, they are provable only under some very strong set-theoretic assumptions going beyond ZFC. Moreover, in contrast to all the above examples of 'mathematical incompleteness', Friedman gave examples of natural ZFC-independent statements of complexity Π_1 in the arithmetical hierarchy. These examples also have only been announced so far ([68], nos. 49–51).

9. Appendix: Relative interpretations

We first give a definition of interpretation of a model M of signature Ω in another model N of signature Σ. Then we define interpretations between theories. Our definition of interpretation is rather general and admits parameters, an interpretation of objects $x \in M$ by tuples of objects $\vec{x} \in N^k$, and an interpretation of equality in M by a congruence relation.

We say that a *translation* I of the signature Ω into the signature Σ is given if:

1) a formula $D_I(\vec{x}, \vec{p})$ of signature Σ is fixed (the tuples $\vec{x}$ and $\vec{p}$ exhaust all free variables of the formula D_I and can be of different lengths k and m, respectively);

2) to any symbol of Ω a formula of signature Σ with the corresponding arity is assigned,

$$\begin{aligned} P &\mapsto P_I(\vec{x}_1, \dots, \vec{x}_n, \vec{p}), \\ f &\mapsto F_I(\vec{x}_1, \dots, \vec{x}_n, \vec{y}, \vec{p}), \\ c &\mapsto C_I(\vec{x}, \vec{p}), \end{aligned}$$

where P and f are an n-place predicate symbol and a function symbol, respectively, and c is a constant in Ω. (All tuples $\vec{x}_i$ of variables are of length k and those of the form $\vec{p}$ are of length m.) In particular, corresponding to the equality symbol in Ω is a formula $=_I (\vec{x}, \vec{y}, \vec{p})$ of signature Σ.

Consider an arbitrary model $(N; \vec{e})$ of signature Σ with a distinguished tuple $\vec{e}$ of constants. The translation I defines in N^k the set

$$M_I = \{\vec{a} \in N^k : N \vDash D_I[\vec{a}, \vec{e}]\},$$

together with the predicates P_I, F_I, and C_I, and $=_I$ defined on N^k.

Definition 10. For a given $\vec{e}$ a translation I is an *interpretation of M in N* if the following conditions hold:

1) $=_I (\vec{x}, \vec{y}, \vec{e})$ satisfies in M_I the equality axioms for the signature Ω, that is, defines a congruence relation on M_I;

2) the predicates F_I and C_I define on the set M_I, modulo $=_I$, a function f_I and a constant c_I, respectively;

3) the model $(M_I; P_I, f_I, c_I)/ =_I$ is isomorphic to M.

A tuple $\vec{e}$ of constants for which these conditions are satisfied is said to be *admissible* for the given interpretation I. The interpretation I has *definable parameters* if for some formula $\mathrm{Par}_I(\vec{p})$ of signature Σ we have

1) $N \vDash \exists \vec{x}\ \mathrm{Par}_I(\vec{x}\,)$,

2) if $N \vDash \mathrm{Par}_I[\vec{e}]$, then the tuple $\vec{e}$ is admissible for I.

A formula A is said to be *simplified* if every atomic subformula occurring in A is of the form $P(x_1, \dots, x_n)$, $f(x_1, \dots, x_n) = y$, or $c = x$, where $x, x_1, \dots, x_n, y$ are variables, P and f are a predicate symbol and a function symbol of the signature, and c is a constant. As is well known, every formula is equivalent to some simplified formula in first-order logic.

Let I be a translation of the signature Ω into Σ. To each variable x (of the language of Ω) we assign its own k-tuple $\vec{x}$ of variables (of the language of Σ). We define a translation $A^I(\vec{x}_1, \dots, \vec{x}_n, \vec{p}\,)$ of a simplified formula $A(x_1, \dots, x_n)$ of signature Ω by induction on the construction of A:

(i) $P(x_1,\dots,x_n)^I \overset{\text{def}}{\Longleftrightarrow} P_I(\vec{x}_1,\dots,\vec{x}_n,\vec{p})$,
$(c=x)^I \overset{\text{def}}{\Longleftrightarrow} C_I(\vec{x},\vec{p})$ and $f(x_1,\dots,x_n)=y \overset{\text{def}}{\Longleftrightarrow} F_I(\vec{x}_1,\dots,\vec{x}_n,\vec{y},\vec{p})$;

(ii) $(\neg A)^I \overset{\text{def}}{\Longleftrightarrow} \neg A^I$, $(A\wedge B)^I \overset{\text{def}}{\Longleftrightarrow} (A^I\wedge B^I)$,

(iii) $\big(\forall x\, A(x)\big)^I \overset{\text{def}}{\Longleftrightarrow} \forall\vec{x}\,\big(D_I(\vec{x},\vec{p})\to A^I(\vec{x},\vec{p})\big)$,

(iv) $\big(\exists x\, A(x)\big)^I \overset{\text{def}}{\Longleftrightarrow} \exists\vec{x}\,\big(D_I(\vec{x},\vec{p})\wedge A^I(\vec{x},\vec{p})\big)$.

By the translation of an arbitrary formula A of signature Ω we mean the translation of a simplified formula equivalent to the given one. This simplified formula is uniquely defined up to logical equivalence.

Let $I(a)$ be an element of the set $M_I\subseteq N^k$ and let $I(a)$ correspond to $a\in M$ under the interpretation I of the model M in N. By induction on the construction of A, we obtain the following proposition.

Proposition 4. *For any formula A of signature Ω, any admissible $\vec{e}\in N$, and any $a_1,\dots,a_n\in M$,*

$$M\vDash A[a_1,\dots,a_n] \iff N\vDash A^I[I(a_1),\dots,I(a_n),\vec{e}].$$

We denote by $\mathrm{Th}(N)$ the set of all sentences that are true in the model M.

Corollary 6. *If M is interpretable in N with definable parameters and if the elementary theory $\mathrm{Th}(N)$ is decidable, then so is $\mathrm{Th}(M)$.*

Proof. For any sentence A in the language of M we have

$$M\vDash A \iff N\vDash \forall\vec{p}\,\big(\mathrm{Par}_I(\vec{p})\to A^I(\vec{p})\big).$$

Thus, to verify the validity of A in M, it suffices to verify the validity of the formula $\forall\vec{p}\,\big(\mathrm{Par}_I(\vec{p})\to A^I(\vec{p})\big)$ in the model N.

Let a theory T of signature Ω and a theory U of signature Σ be given. We assume that T is defined by a set of axioms which are closed formulae.

Definition 11. A translation I with definable parameters is called an *interpretation of the theory T in U* if

1) $U\vdash \mathrm{Par}_I(\vec{p})\to A^I(\vec{p})$ for any axiom $A\in T$;

2) $U\vdash \mathrm{Par}_I(\vec{p})\to \exists\vec{x}\, D_I(\vec{x},\vec{p})$;

3) $U\vdash \mathrm{Par}_I(\vec{p})\to E^I(\vec{p})$, where E is the equality axiom for the signature Ω;

4) $U\vdash \mathrm{Par}_I(\vec{p})\to \big(\forall x_1\dots x_n\,\exists! y\, f(x_1,\dots,x_n)=y\big)^I$ for the function symbols f in Ω;

5) $U\vdash \mathrm{Par}_I(\vec{p})\to (\exists! x\, c=x)^I$ for the constants c in Ω.

The theory T is said to be *interpretable in U* if there is an interpretation of T in U.

Thus, in any model N of the theory U and for a choice of parameters satisfying the condition Par_I, the translation I defines a model of the theory T.

By induction on the length of the derivation of a formula A we obtain the following proposition.

Proposition 5. *If I is an interpretation of T in U and if $T \vdash A$, then*

$$U \vdash \forall \vec{x}\,\big(\mathrm{Par}_I(\vec{p}) \wedge D_I(\vec{x},\vec{p}) \;\to\; A^I(\vec{x},\vec{p})\big).$$

Corollary 7. *If T is interpretable in U and U is consistent, then so is T.*

This result is established by elementary (syntactic) methods not based on set theory. As a rule, consistency proofs based on the existence of a model go beyond the framework of elementary methods, because the models are usually infinite and are constructed in the framework of set theory. The method of interpretations enables one to avoid unnecessary hypotheses about the existence of infinite sets and leads to a 'finitary' reduction of one theory to another. Using this approach, one can prove, in particular, the consistency of the Lobachevskii elementary geometry with respect to elementary Euclidean geometry, the consistency of the continuum hypothesis (and the consistency of its negation) with respect to the set theory ZFC, and other important results.

Bibliography

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

[2] R. Zach, "Hilbert's Program", *The Stanford Encyclopedia of Philosophy* (E. N. Zalta, ed.) 2009, http://plato.stanford.edu/archives/spr2009/entries/hilbert-program/.

[3] C. Smorinskii, "The incompleteness theorems", *Handbook of mathematical logic* (J. Barwise, ed.), Stud. Logic Found. Math., vol. 90, North-Holland, Amsterdam–New York–Oxford 1977, pp. 821–865.

[4] S. C. Kleene, "Introductory note to 1930b, 1931 and 1932b", *Kurt Gödel. Collected Works* (S. Feferman et al., ed.), vol. 1: *Publications* 1929–1936, The Clarendon Press, Oxford Univ. Press, New York 1986, pp. 126–141.

[5] K. Gödel, "On undecidable propositions of formal mathematical systems", *Kurt Gödel. Collected Works*, vol. 1: *Publications* 1929–1936, The Clarendon Press, Oxford Univ. Press, New York 1986, pp. 346–373.

[6] S. Feferman, J. R. Dawson, S. C. Kleene, G. H. Moore, R. M. Solovay, and J. van Heijenoort (eds.), *Kurt Gödel. Collected Works*, vol. 1: *Publications* 1929–1936, The Clarendon Press, Oxford Univ. Press, New York 1986.

[7] A. Tarski, "Der Wahrheitsbegriff in der formalisierten Sprachen", *Stud. Philos.* **1** (1935), 261–405.

[8] A. Ehrenfeucht and S. Feferman, "Representability of recursively enumerable sets in formal theories", *Arch. Math. Logik Grundlag.* **5**:1-2 (1960), 37–41.

[9] R. M. Smullyan, *Diagonalization and self-reference*, Oxford Logic Guides, vol. 27, The Clarendon Press, Oxford Univ. Press, New York 1994, 396 pp.

[10] J. B. Rosser, "Gödel theorems for non-constructive logics", *J. Symbolic Logic* **2**:3 (1937), 129–137.

[11] W. Craig, "On axiomatizability within a system", *J. Symbolic Logic* **18**:1 (1953), 30–32.

[12] S. C. Kleene, "General recursive functions of natural numbers", *Math. Ann.* **112**:1 (1936), 727–742.

[13] S. C. Kleene, "Recursive predicates and quantifiers", *Trans. Amer. Math. Soc.* **53**:1 (1943), 41–73.

[14] S. C. Kleene, "A symmetric form of Gödel's theorem", *Indag. Math.* **12** (1950), 244–246.

[15] S. C. Kleene, *Introduction to metamathematics*, Van Nostrand, New York 1952, 550 pp.

[16] R. M. Smullyan, *Theory of formal systems*, Ann. Math. Stud., vol. 47, Princeton Univ. Press, Princeton, NJ 1961, 142 pp.

[17] В. А. Успенский, "Теорема Гёделя о неполноте в элементарном изложении", *УМН* **29**:1 (1974), 3–47; English transl., V. A. Uspenskii, "An elementary exposition of Gödel's incompleteness theorem", *Russian Math. Surveys* **29**:1 (1974), 63–106.

[18] B. Rosser, "Extensions of some theorems of Gödel and Church", *J. Symbolic Logic* **1**:3 (1936), 87–91.

[19] A. Mostowski, "On definable sets of positive integers", *Fund. Math.* **34** (1947), 81–112.

[20] W. V. Quine, "Concatenation as a basis for arithmetic", *J. Symbolic Logic* **11**:4 (1946), 105–114.

[21] A. Tarski, A. Mostowski, and R. M. Robinson, *Undecidable theories*, Stud. Logic Found. Math., North-Holland, Amsterdam 1953, 98 pp.

[22] A. S. Troelstra and H. Schwichtenberg, *Basic proof theory*, Cambridge Tracts Theoret. Comput. Sci., vol. 43, Cambridge Univ. Press, Cambridge 1996, 343 pp.

[23] J. R. Shoenfield, *Mathematical logic*, Addison-Wesley, Reading, MA–London–Don Mills, ON 1967, 344 pp.

[24] P. Pudlák, "Cuts, consistency statements and interpretations", *J. Symbolic Logic* **50**:2 (1985), 423–441.

[25] S. Feferman, "Arithmetization of metamathematics in a general setting", *Fund. Math.* **49** (1960), 35–92.

[26] P. Hájek and P. Pudlák, *Metamathematics of first-order arithmetic*, Perspect. Math. Logic, Springer-Verlag, Berlin 1993, 460 pp.

[27] Ю. В. Матиясевич, "Диофантовость перечислимых множеств", *Докл. АН СССР* **191**:2 (1970), 279–282; English transl., Yu. V. Matiyasevich, "Enumerable sets are Diophantine", *Soviet Math. Dokl.* **11** (1970), 354–358.

[28] Ю. В. Матиясевич, *Десятая проблема Гильберта*, Наука, М. 1993, 224 pp.; English transl., Yu. V. Matiyasevich, *Hilbert's tenth problem*, Found. Comput. Ser., MIT Press, Cambridge, MA 1993, 264 pp.

[29] Ю. Л. Ершов, *Проблемы разрешимости и конструктивные модели*, Наука, М. 1980, 416 с. [Yu. L. Ershov, *Decision problems and constructivizable models*, Nauka, Moscow 1980, 416 pp.]

[30] Ю. Л. Ершов, И. А. Лавров, А. Д. Тайманов, М. А. Тайцлин, "Элементарные теории", *УМН* **20**:4 (1965), 37–108; English transl., Yu. L. Ershov, I. A. Lavrov, A. D. Taimanov, and M. A. Taitslin, "Elementary theories", *Russian Math. Surveys* **20**:4 (1965), 35–105.

[31] A. Bès, "A survey of arithmetical definability. A tribute to Maurice Boffa", *Bull. Belg. Math. Soc. Simon Stevin*, 2001, suppl., 1–54.

[32] V. Švejdar, "An interpretation of Robinson arithmetic in its Grzegorczyk's weaker variant", *Fund. Inform.* **81**:1–3 (2007), 347–354.

[33] R. L. Vaught, "On a theorem of Cobham concerning undecidable theories", *Logic, Methodology and Philosophy of Science* (Proc. 1960 Internat. Congr.), Stanford Univ. Press, Stanford 1962, pp. 14–25.

[34] V. Švejdar, "Weak theories and essential incompleteness", *The Logica Yearbook* 2007 (M. Peliš, ed.), Proceedings of the Logica 07 International Conference, Philosophia, Prague 2008, pp. 213–224.

[35] A. Visser, "Pairs, sets and sequences in first-order theories", *Arch. Math. Logic* **47**:4 (2008), 299–326.

[36] A. Visser, *Cardinal arithmetic in weak theories*, Logic Group Preprint Series, **265**, Department of Philosophy, Univ. Utrecht 2008, http://igitur-archive.library.uu.nl/lg/2008-0422-200811/UUindex.html.

[37] H. Putnam, "Decidability and essential undecidability", *J. Symbolic Logic* **22**:1 (1957), 39–54.

[38] A. Ehrenfeucht, "Two theories with axioms built by means of pleonasms", *J. Symbolic Logic* **22**:1 (1957), 36–38.

[39] J. P. Jones and J. C. Shepherdson, "Variants of Robinson's essentially undecidable theory R", *Arch. Math. Logik Grundlag.* **23**:1 (1983), 61–64.

[40] A. Visser, *Why the theory R is special*?, Logic Group Preprint Series, **279**, Department of Philosophy, Univ. Utrecht 2009, http://igitur-archive.library.uu.nl/ph/2009-0812-200111/UUindex.html.

[41] R. L. Vaught, "Axiomatizability by a schema", *J. Symbolic Logic* **32**:4 (1967), 473–479.

[42] S. Feferman, "Finitary inductively presented logics", *Logic Colloquium'88* (Padova, 1988), Stud. Logic Found. Math., vol. 127, North-Holland, Amsterdam 1989, pp. 191–220.

[43] A. Grzegorczyk, "Undecidability without arithmetization", *Studia Logica* **79**:2 (2005), 163–230.

[44] F. Ferreira, "Polynomial time computable arithmetic", *Logic and computation* (Pittsburgh, PA, 1987), Contemp. Math., vol. 106, Amer. Math. Soc., Providence, RI 1990, pp. 137–156.

[45] Ю. Л. Ершов, *Определимость и вычислимость*, Сибирская школа алгебры и логики, Научная книга, Новосибирск 1996, 287 с.; English transl., Yu. L. Ershov, *Definability and computability*, Siberian School of Algebra and Logic, Consultants Bureau, New York 1996, 262 pp.

[46] G. J. Chaitin, "Information-theoretic limitations of formal systems", *J. Assoc. Comput. Mach.* **21**:3 (1974), 403–424.

[47] G. Boolos, "A new proof of Gödel's incompleteness theorem", *Notices Amer. Math. Soc.* **36** (1989), 388–390; G. Boolos, "A new proof of Gödel's incompleteness theorem", *Logic, logic, and logic*, Harvard Univ. Press, Cambridge, MA 1998, pp. 383–388.

[48] P. Raatikainen, "On interpreting Chaitin's incompleteness theorem", *J. Philos. Logic* **27**:6 (1998), 569–586.

[49] M. Kikuchi, "Kolmogorov complexity and the second incompleteness theorem", *Arch. Math. Logic* **36**:6 (1997), 437–443.

[50] В. А. Успенский, *Теорема Гёделя о неполноте и четыре дороги, ведущие к ней. Лекция* 1, http://www.mathnet.ru/php/presentation.phtml?option_lang=rus&presentid=122. [V. A. Uspenskii, *Gödel's incompleteness theorem and* 4 *roads to it. Lecture* 1, Lectures of 'Contemporary Mathematics' Summer School, Dubna 2007.]

[51] В. А. Успенский, *Теорема Гёделя о неполноте*, Популярные лекции по математике, Наука, М. 1982, 110 с. [V. A. Uspenskii, *Gödel's incompleteness theorem*, Popular Mathematics Lectures, Nauka, Moscow 1982, 110 pp.]

[52] E. Mendelson, *Introduction to mathematical logic*, Van Nostrand, Princeton, NJ 1964, 300 pp.

[53] V. H. Dyson, J. P. Jones, and J. C. Shepherdson, "Some Diophantine forms of Gödel's theorem", *Arch. Math. Logik Grundlag.* **22**:1-2 (1982), 51–60.

[54] П. С. Новиков, "Об алгоритмической неразрешимости проблемы тождества слов в теории групп", Тр. МИАН СССР, **44**, Изд-во АН СССР, М. 1955, с. 3–143; English transl., P. S. Novikov, *On the algorithmic insolvability of the word problem in group theory*, Amer. Math. Soc. Transl. Ser. 2, vol. 9, Amer. Math. Soc., Providence, RI 1958, 122 pp.

[55] С. И. Адян, "Алгоритмическая неразрешимость проблем распознавания некоторых свойств групп", *Докл. АН СССР* **103**:4 (1955), 533–535. [S. I. Adian, "Algorithmic unsolvability of the problem of recognition of some group properties", *Dokl. Akad. Nauk SSSR* **103**:4 (1955), 533–535.]

[56] С. И. Адян, "Неразрешимость некоторых алгоритмических проблем теории групп", Тр. ММО, **6**, 1957, с. 231–298. [S. I. Adian, "Unsolvability of several algorithmic problems in group theory", Tr. Mos. Mat. Obshch., vol. 6, 1957, pp. 231–298.]

[57] A. Bovykin, "Brief introduction to unprovability", *Logic Colloquium* 2006 (Radboud University, Nijmegen, 2006), Proceedings of Annual European Conference on Logic of the Association for Symbolic Logic, Lect. Notes Log., Assoc. Symbol. Logic, Chicago, IL; Cambridge Univ. Press, Cambridge Univ. Press 2009, pp. 38–64.

[58] J. Paris and L. Harrington, "A mathematical incompleteness in Peano arithmetic", *Handbook of mathematical logic* (J. Barwise, ed.), Stud. Logic Found. Math., vol. 90, North-Holland, Amsterdam–New York–Oxford 1977, pp. 1133–1142.

[59] L. A. S. Kirby and J. Paris, "Accessible independence results for Peano arithmetic", *Bull. London Math. Soc.* **14**:4 (1982), 285–293.

[60] A. Kanamori and K. McAloon, "On Gödel's incompleteness and finite combinatorics", *Ann. Pure Appl. Logic* **33**:1 (1987), 23–41.

[61] L. D. Beklemishev, "The Worm principle", *Logic Colloquium'*02 (Munster, 2002), Joint proceedings of the Annual European Summer Meeting of the Association for Symbolic Logic and the Biannual Meeting of the German Association for Mathematical Logic and the Foundations of Exact Sciences (the Colloquium Logicum), Lect. Notes Log., vol. 27, Assoc. Symbol. Logic, La Jolla, CA; A K Peters, Ltd., Wellesley, MA 2006, pp. 75–95; Logic Group Preprint Series, **219**, Utrecht Univ. March, 2003.

[62] M. Hamano and M. Okada, "A relationship among Gentzen's proof-reduction, Kirby–Paris' hydra game, and Buchholz's hydra game", *Math. Logic Quart.* **43**:1 (1997), 103–120.

[63] S. G. Simpson, "Nichtbeweisbarkeit von gewissen kombintorischen Eigenschaften endlicher Bäume", *Arch. Math. Logik Grundlag.* **25**:1 (1985), 45–65.

[64] M. Rathjen and A. Weiermann, "Proof-theoretic investigations on Kruskal's theorem", *Ann. Pure Appl. Logic* **60**:1 (1993), 49–88.

[65] H. M. Friedman, "Internal finite tree embeddings", *Reflections on the foundations of mathematics* (Stanford, CA, 1998), Lect. Notes Log., vol. 15, Assoc. Symbol. Logic, Urbana, IL; A K Peters, Ltd., Natick, MA 2002, pp. 60–91.

[66] W. Buchholz, "An independence result for $(\Pi^1_1\text{-CA}) + \text{BI}$", *Ann. Pure Appl. Logic* **33**:2 (1987), 131–155.

[67] H. M. Friedman, "Finite functions and the necessary use of large cardinals", *Ann. of Math.* (2) **148**:3 (1998), 803–893, arXiv: math/9811187.

[68] H. M. Friedman, Downloadable manuscripts at `http://www.math.ohio-state.edu/~friedman/manuscripts.html`.

L. D. Beklemishev
Steklov Mathematical Institute, Russian Academy of Sciences
E-mail: `bekl@mi.ras.ru`

Received 20/AUG/10
Translated by A. SHTERN