On the existence of absolutely undecidable sentences of number theory
and convincing unformalizable mathematical arguments

Abstract

This paper deals with the concept of absolute undecidability in mathematics. The main technical result concerns a particular sentence $D$ expressed in the language of Peano Arithmetic. On basis of this result, it is argued that it is not possible to decide $D$ in any “reasonable” formal theory $T$, under a wide definition of “reasonable”. The undecidability argument is independent of both $T$ and its language. Therefore, it can be viewed as an argument for the absolute undecidability of the sentence $D$. While the proof of the technical result itself is a standard, formalizable proof, the associated undecidability argument concerning $D$ can be viewed as an example of a valid mathematical argument which by nature is not formalizable. Hence, as a by-product, the construction is relevant for the general discussion about the existence of such arguments.

1. Introduction

This paper adresses the notion of absolute undecidability in mathematics. An overview over relevant historical material concerning this can be found in Koellner [2006]. See also Feferman [2005].

As is pointed out in Koellner [2006], the notion of absolute undecidability is very different from undecidability relative to a given theory. Examples of the latter kind may be produced by applying Gödel’s incompleteness theorems. Also in this category, we have independence results such as the independece of the continuum hypothesis (CH) in Zermelo-Fraenkel set theory (ZF). While the Gödel results show that for each given (axiomatizable) theory there are “true” statements which cannot be proved in the theory, there remains the possibility that each of these statements may be proved in another, more powerful theory in the same language, with the same intended model as the original one. In contrast, by an absolutely undecidable mathematical sentence we mean a statement that is, to quote Koellner [2006], “undecidable relative to any set of axioms which are justified”.

In Koellner [2006], three candidates for absolutely undecidable sentences which are all expressible in the language of the theory ZF, are discussed. The continuum hypothesis is one of these. Koellner concludes that there are promising approaches to obtaining reasonable axioms that could potentially settle a “truth value” for all the three candidates, and finishes his paper with the following two sentences (quote): “There is at present no solid argument to the effect that a given statement is absolutely undecidable. We do not even have a clear scenario for how such an argument might go.” (Koellner [2006])

The last sentence puts the finger on exactly what is the problem if one tries to argue for “absolute undecidability” of, say, the sentence CH. There is no really intrinsic property of CH which indicates that is absolutely undecidable. On the contrary, it is a relatively simple and neat statement expressible in the language of ZF. For this reason, to argue for its “absolute undecidability” one has to somehow prove that no one will ever come up with some reasonably self-evident property of sets
which can be shown to imply CH. Part of the problem here is that the language of set theory is very powerful, and that the notion of a “set”, even when described by the most powerful “reasonable” axioms systems we have, is still a bit too vague.

In this paper, I will construct another candidate for an absolutely undecidable sentence; in fact I will present an explicit, theory-independent argument for its absolute undecidability. This is the sentence $D$ described by the theorem in section 5. The sentence $D$ is expressed in the simple language $L(\text{PA})$ of Peano Arithmetic (PA). Unlike CH, this sentence will have an intrinsic property which can be used to argue for its absolute undecidability. This property will be one that can be perceived as “infinite complexity” from a certain point of view. However, it must be emphasized that of course I am talking about undecidability regardless of physical limitations in time and space here. I am not talking about the kind of “undecidability” enjoyed by a sentence which is so complex that we will never be able to decide it for physical reasons. That kind of undecidability certainly does not deserve the label “absolute”. The “infinite” complexity buried in the formula $D$ derives from a certain way of going through all possible extensions of PA (regardless of the language), collecting some information which is inaccessible to each extension, and combining the information by using, among all things, an infinite product of rational numbers in the interval $(0, 1)$ converging to a real number in $(0, 1)$.

Now many well-known theories with “reasonable” axioms, notably Zermelo-Fraenkel set theory, actually prove the existence of a definite truth value in the standard model $\mathbb{N}$ of natural numbers for every sentence of the language $L(\text{PA})$. For this reason, one might guess that trying to find an absolutely undecidable sentence expressible in $L(\text{PA})$, is hopeless. Thus one turns instead to more powerful languages, such as the language of ZF. However, these things are not as simple as they may seem at first sight. Although we may decide to accept as “true” the axioms of ZF, and ZF proves the existence of a truth value “true” or “false” for every $L(\text{PA})$ sentence, this does not rule out the existence of $L(\text{PA})$ sentences whose truth values are “nonconstructible” in the sense that we can never find them. The fact that a complete definition of truth for $L(\text{PA})$ sentences in $\mathbb{N}$ can be consistently assumed and also proved as a raw existence assertion in a theory with reasonable axioms, does not imply that there actually is such a complete concept of truth in the structure $\mathbb{N}$ built on the concept of “finite” natural numbers. It could happen that unbounded complexity or the non-formalizable nature of the finite/infinite distinction plays us a trick which is rendering truth definitions by induction on the complexity of formulas incomplete. Thus the material in this paper does not, at least as far as I can tell, link to inconsistency questions concerning theories such as Zermelo-Fraenkel set theory, in which completeness of the truth concept in $\mathbb{N}$ can be proved. The important point is that when we are dealing with absolute undecidability, the issue is underdetermination, not overdetermination. This will become clearer in the following, and it leads us from the main question addressed in this paper, namely the existence of “absolutely undecidable” mathematical statements, over to the question of whether or not there exist mathematically valid forms of reasoning which are not formalizable.

Making this paper readable for its intended audience I have found challenging. The main issues addressed are philosophical, and therefore the paper should be made readable to people who are interested in the philosophy of mathematics, but who are not necessarily logicians. On the other hand, some technicalities are needed in order to derive and explain the results, even on an intuitive level. My solution has become to mix elementary logical theory and intuitive explanations with the more technical material needed, thereby hopefully making the paper readable at different technical
levels. As my reference for standard mathematical logic, I use the classic book Shoenfield [1967]. When I refer to “standard theory”, this reference may always be used. Hopefully, my philosophically inclined reader will be able to figure out some of the necessary technicalities directly and fill in the rest by consulting references such as Shoenfield [1967].

2. Extensions of Peano Arithmetic

We take our departure in the simple formal language $L = L(\text{PA})$ of Peano Arithmetic (PA). This language may be interpreted in terms of the natural numbers 0, 1, 2, 3, .... The language may be defined so that it has the symbols

$$\forall, \exists, \neg, \land, \lor, \leftrightarrow, (, ), =, +, -, >, O, S, x, y, z, w, \prime$$

It is customary to exclude the symbols $\forall, \land, \lor, \leftrightarrow$ and $>$ from this list, and define them instead as abbreviations for certain combinations of the remaining symbols. Even the left and right parenthesis signs may be eliminated. For our purposes however, one may just as well work directly with all the symbols listed above.

In the standard structure $N$ of $L$, which may be thought of as the intended interpretation of $L(\text{PA})$, the first 15 of the symbols listed are interpreted, respectively, as “for all”, “there exists”, negation, or, and, implies, if and only if (iff), left and right parentheses, equality, plus, multiply, smaller than, bigger than, and zero. The symbol $S$ is interpreted as “successor”, so that SO is interpreted as 1, SSO as 2, and so on. The letters $x, y, z, w$ are used as variables, and the $\prime$ is used in cases where more than four variables are needed. Thus we may form variables $x', x'', x'''$ and so on. An example of a formula in the language $L$ of PA is

$$\forall x \exists y (y > Sx)$$

This states that for all $x$, there exists $y$ such that $y > x + 1$. Interpreted in terms of natural numbers, this is a true statement. The formula (1) has no free variables, that is, all the variables are quantified by “for all” or “there exists”. Such formulas are called sentences.

In the following, when I use words like “formula” and “sentence” without specifying the language, it is always understood that the language in question is $L$.

The theory PA itself is expressed in the language $L$. We may say that the language of PA is $L$, or that PA is an $L$ theory. The theory PA has a set of axioms which are formulas of $L$. These are translations into $L$ of statements which are generally considered as “self-evident” statements about natural numbers.

Also, PA has some rules of inference which can be used to derive results (theorems) of the theory. These rules of inference, which are common for all so-called first order logical theories, are such that correctness of proofs in PA may be checked algorithmically. By replacing the axioms of PA with other formulas of $L$, we obtain other $L$ theories.

In general, a model of a logical theory $T$ is a structure (“interpretation”) for the language of $T$ which is such that all the axioms are rendered true. The standard model of PA, which may be thought of simply as “the natural numbers”, is $N$.

By standard theory, we may choose a Gödel numbering to code formulas in the language $L = L(\text{PA})$ into natural numbers. We may also use our Gödel numbering of formulas to code finite, ordered sequences of formulas, by employing sequence numbers in the usual way. If $A$ is
a formula of $L$, then the Gödel number of $A$ is denoted $[A]$. I will use the term “Gödel number” also for sequence numbers defined for sequences of formulas, based on the Gödel numbering for formulas. In particular, I will speak of “the Gödel number” of a proof in a given theory.

We are now going to focus on a particular set of formulas in $L(PA)$, namely the set of $\Delta_0$ formulas $A(x, y)$ with the two free variables $x$ and $y$. Roughly speaking, a $\Delta_0$ formula is a formula which has no unbounded quantifiers. More precisely, suppose that $A(x, y)$ is a $\Delta_0$ formula of $L(PA)$ with the two free variables $x$ and $y$. Let $n$ and $m$ are arbitrary natural numbers. Then the concrete instance $A(n, m)$ may be rewritten as a formula with no quantifiers ($\exists$ and $\forall$). Here and in the following, $n$ denotes the numerical term $SSS...O$ with $n$ copies of $S$, corresponding to the natural number $n$. Another way of expressing the $\Delta_1^0$ condition on $A(x, y)$ is to say that all quantifiers occurring in the formula $A(x, y)$ are bounded in the sense that each quantifier is accompanied by a condition of the type $v<k$ on the quantified variable $v$, where the upper bounds $k$ may depend on the values of $x$ and $y$, but where the dependences are of such a nature that if the values of $x$ and $y$ are known, then quantifier elimination may be achieved by rewriting the formula.

Now let $E^+ \subseteq N$ be the set of Gödel numbers of $\Delta_0$ formulas $A(x, y)$ of $L(PA)$ with the two free variables $x$ and $y$. By standard theory, we may choose a $\Delta_0$ formula $e^+(z)$ in $L$ with the single variable $z$ which defines the set $E^+$. To put it another way, the formula $e^+(z)$ defines the property of being the Gödel number of a $\Delta_0$ formula with the two free variables $x$ and $y$ in the $L$-structure $N$. Expressed in a more technical fashion, this means that for all natural number $n$, we have

$$n \in E^+ \iff N \models e^+(n).$$

It is important to emphasize that both here and in the following, when I speak of “choosing” a formula, I do not refer to defining the formula by, say, some $L$-expressible properties which could, potentially, behave differently in different models of PA. I am simply talking about choosing a concrete formula of $L$, a formula which we could in principle construct explicitly, and keeping that formula fixed in our subsequent theoretical development.

Concerning the formula $e^+(z)$, it follows from standard theory that we may also assume, as a basis for our choice, that the formula not only defines the property of being in the set $E^+$, but also that $e^+(z)$ represents this property with respect to PA. This means that PA is able to decide (prove or disprove) any numerical instance $e^+(n)$ of $e^+(z)$.

Among the formulas in $E^+$, by standard theory we may choose a formula $Pr_{PA}(x, y)$ arithmetizing the proof relation in PA with respect to PA itself. Then if $n$ is the Gödel number of an $L(PA)$ formula $B$ and $m$ is the Gödel number of a proof of $B$ in PA, we have

$$PA \vdash Pr_{PA}(n, m)$$

Otherwise, we have

$$PA \vdash \neg Pr_{PA}(n, m).$$

The numerical term corresponding to the Gödel number of $Pr_{PA}(x, y)$ will be denoted by $pa$.

In a similar fashion, some of the formulas $A(x, y)$ coded by other numbers in our set $E^+$ may be thought of as arithmetizing the proof relation in other “formal” theories $T$. This produces a very general setup, as will be explained further in section 6. So far, the requirement on $T$ is simply that there is a $\Delta_0$ formula in $L(PA)$ arithmetizing the proof relation in $T$, by means of some Gödel
numbering of the language $L(T)$ of $T$. The language $L(T)$ does not need to be $L(PA)$, nor does $T$ need to be a first order logical theory. In section 6, we will interpret things in such a way that if

$$A(n, m)$$

is true in $N$, then the natural number $n$ is the $L(T)$ Gödel number of an $L(T)$ formula $B$, and the natural number $m$ is the Gödel number of a proof in $T$ of $B$. In this situation, we may say that

$$\exists y A(\lceil B \rceil, y)$$

means that $T$ proves $B$, which we write $T \vdash B$.

Now order the set $E^+$ by increasing size, and let $E^+_n$ be the $n$th number in the sequence, for $n = 1, 2, 3, \ldots$. Since the set $E^+$ is represented in $PA$ by the formula $e^+(z)$, it follows from standard theory that we may also choose, or explicitly construct if you will, an $L$-formula

$$\text{ORD}e^+(w, x)$$

with two free variables representing the relation

$$w = E^+_x$$

with respect to $PA$. The fact that $PA$ is able to decide (prove or disprove) all numerical instances of the formula $\text{ORD}e^+(w, x)$, will play a crucial role in the proof of the theorem in section 5.

While the set $E^+$ is wide enough to cover provability in all the formal theories we are interested in, not all the $L(PA)$ formulas with Gödel numbers in $E^+$ can be meaningfully thought of as representing a provability condition. For this reason, we are going to define a subset $E \subseteq E^+$ which we will officially label as the set of extensions of $PA$. While the set $E$ will still be wide enough for containing the Gödel numbers of all $L$-formulas $A(x, y)$ coding provability in reasonable “formal” mathematical theories which have at least the strength of $PA$ when it comes to deriving facts about natural numbers, the additional conditions imposed on it will be sufficient for making our interpretations in section 6 work out. Before we describe the conditions defining $E$, it is convenient to introduce some abbreviations for formulas.

First of all, we want $z \vdash w$ to be an abbreviation for an $L(PA)$ formula expressing that $z \in E^+$, and that the formula $A(x, y)$ with Gödel number $z = \lceil A(x, y) \rceil$ “proves” the $L$-formula with Gödel number $w$, in the sense that

$$\exists y A(w, y)$$

is true in $N$. Technically, this may be achieved by choosing $z \vdash w$ to be the formula

$$e^+(z) \land \exists y \text{Sat}_{\Sigma_1}(\lceil z[w, y] \rceil),$$

where $\text{Sat}_{\Sigma_1}(x)$ is a $\Sigma_1$ truth predicate formula for $\Sigma_1$ sentences of $L$, and where $\lceil z[w, y] \rceil$ means the Gödel number of the formula obtained by substituting $w$ for $x$ and $y$ for $y$ in the formula with Gödel number equal to the value of $z$. We choose such a truth predicate and keep it fixed in the following. An explicit construction of $\text{Sat}_{\Sigma_1}(x)$ can be found in chapter 9 of Kaye (1991). If $n$ and $m$ are natural numbers, then we say that $n$ proves $m$, and sometimes write $n \vdash m$, iff the formula

$$n \vdash m$$
is true in $N$. We may then also write $n \vdash m$ or $n \vdash m$. We say that $n$ decides $m$ iff $n$ proves $m$ or $n$ proves the Gödel number of the negation of the formula with Gödel number $m$. We will say that a given number $i \in E^+$ has $N$ as a model part if for all sentences $B$ of $L$ such that $i \vdash \lceil B \rceil$, we have that $B$ is true in $N$.

Next, we need a formula $\text{Con}_z$ expressing that the natural number $z$ codes a consistent theory. A consistent theory may, in our context, be defined as a theory in which the statement $0 = 1$ is not provable. We may let $\text{Con}_z$ be an abbreviation for the following formula with one free variable $z$:

$$e^+(z) \land \neg (z \vdash \lceil 0 = 1 \rceil).$$

Our next task is to introduce a formula $\text{NinCand}(z, w)$ which will be used to describe candidates for a “naturally inaccessible number” (NIN) which we will associate to each theory we consider. Roughly, the idea is that if $z \in E^+$, then $\text{NinCand}(z, w)$ should hold in $N$ iff $w$ represents a formula which is not provable by $z$, and which is also not directly verifiable as such by $z$. To this end, let $\text{NinCand}(z, w)$ be an abbreviation for the following formula with free variables $z$ and $w$:

$$e^+(z) \land \neg (z \vdash w) \land \neg (z \vdash \lceil \text{Con}_z \rightarrow \neg (z \vdash w) \rceil).$$

Roughly described, $\text{NinCand}(z, w)$ states that $z$ should be in $E^+$, that $z$ should not prove $w$, and that $z$ should not be able to prove that it does not prove $w$, even assuming its own consistency.

We are now in a position where we can choose an $L(\text{PA})$ formula $e(z)$ stating the conditions we will use to define our subset $E \subseteq E^+$ of PA extensions. There are four requirements we want to make for a number $r$ to be in $E$: (1) It should be in $E^+$, (2) it should represent a theory proving everything provable in PA, (3) it should represent a theory not proving its own inconsistency, and (4) it should have a NIN candidate. We may then choose $e(r)$ to be the following $L(\text{PA})$ formula with the single free variable $z$:

$$e^+(z) \land \forall w (\text{pa} \vdash w \rightarrow z \vdash w) \land \neg (z \vdash \lceil \neg \text{Con}_z \rceil) \land \exists w \text{NinCand}(z, w).$$

We define $E \subseteq E^+$ as the set of all $a \in E^+$ such that $e(a)$ is true in $N$.

Finally, we will introduce a formula

$$\text{nin}(z, w)$$

which holds in $N$ iff $w$ represents the minimal NIN candidate for $z \in E$, with numbers in $E^+ \setminus E$ having their “minimal NIN candidates” defined equal to 0. Let $\text{nin}(z, w)$ be the $L(\text{PA})$ formula

$$e^+(z) \land ((\neg e(z) \land w = 0) \lor (e(z) \land \text{NinCand}(z, w) \land \neg \exists y (y < w \land \text{NinCand}(z, y)))).$$

It follows that for each natural number $n \in E^+$ there is a unique natural number $m$ such that $\text{nin}(n, m)$ is true in $N$. This natural number $m$ will be called the naturally inaccessible number (NIN) of $n$. Of course, the NIN of $n$ is not really “inaccessible” if $n \notin E$.

As has already been mentioned, the conditions we have put on $E$ will come into play in the undecidability argument of section 6. The idea behind the conditions is to exclude elements in $E^+$.
which could disturb the interpretation we will make of the main theorem. The main theorem is found in section 5. Note that while the set $E^+$ is represented in PA by the formula $e^+(z)$ and its ordering is represented in PA by the formula $\text{ORD}e^+(w, x)$, the set $E \subseteq E^+$ is merely defined by $e(z)$ in $L(\text{PA})$ with respect to truth in the standard model $N$.

3. Contribution sequences

The purpose of this section is to introduce a mapping taking each natural number $a$ into a contribution sequence of binary bits in such a way that these bits are, roughly speaking, indeterminable unless one has substantial information concerning $a$. In the following, we use a more or less arbitrarily chosen example of such a mapping. This choice will be sufficient for our purposes.

Our choice of mapping defining the contribution sequences may be defined as follows. Let $f$ be the function mapping each natural number $a$ to the smallest prime number greater than $a!$. Let an ordered pair of natural numbers $(a, n)$ be given, with $a \neq 0$. Let $q$ be the smallest natural number $v$ such that $v$ may be obtained by applying $f$ to $a$ at least one time, and $\log_2(v) > nn$.

If the number of binary digits in $q$ is $z$, then let $k$ be the smallest digit position bigger than $z/2$. Then for each prime number $p_i$ ordered by increasing size, let

$$(b_{n}^{a})_i$$

be the $q$ bit at position $k + p_i$, for $i = 1, 2, 3, \ldots, n$. The sequence

$$(b_{n}^{a})_1, \ldots, (b_{n}^{a})_n$$

will henceforth be called the length $n$ contribution sequence of the number $a \neq 0$, or the length $n$ contribution sequence based on the number $a$. To make the definition complete, we define the length $n$ contribution sequence of 0 to be the $n$-bit sequence consisting of zeros only. By standard theory, we may choose a formula Contr$(x, y, z, w)$ in $L$ such that

Contr$(a, n, m, k)$

is true in $N$ iff the $m$th bit in the length $n$ contribution sequence based on $a$ has value $k$, and such that Contr$(x, y, z, w)$ represents this relation with respect to PA. We keep the formula Contr$(x, y, z, w)$ fixed in the following.

4. Infinite bit sequences. The sequence $\tau$.

As is well known, infinite bit sequences may be considered as real numbers in the open interval $(0, 1)$ written in base 2 notation. For example,

$$r = 1010111000111\ldots$$ may be identified with \hspace{1cm} $0.1010111000111\ldots$ \hspace{1cm} (2)

The set of all infinite bit sequences is often called Cantor space. In Cantor space there is a probability measure which is sometimes called the fair coin measure. This measure is defined so that if $w$ is
a finite binary string, then the probability measure of the set of all infinite sequences which begin with \( w \), is

\[ p_w = \left( \frac{1}{2} \right)^{\text{number of bits in } w} \]

We can interpret this by saying that the *property of having \( w \) as an initial segment*, has probability \( p_w \) in Cantor space. In general, if \( P \) is some property of infinite bit sequences, and the set of all infinite bit sequences with this property has probability measure \( p \), then we say that \( P \) is a *probability p property* or a *\( p \) probability property* of infinite bit sequences.

As an example, let \( r \) be the infinite bit sequence \( (2) \). If \( P \) is the property of “agreeing with \( r \) on the first three bits”, then the probability of \( P \) is \( (1/2)^3 = 1/8 \). Then \( P \) is a probability \( 1/8 \) property. We may interpret this result by saying that if the bits of a sequence in Cantor space are obtained successively by flipping a fair coin, then the probability of ending up with a sequence having our property \( P \), is \( 1/8 \).

I will give another example. Imagine once again that the bits of an infinite bit sequence \( r \) are chosen successively at random, with probability 0.5 for 0 and 1 at each bit position. Then the probability for the event that we will end up with a sequence \( r \) with *only a finite numbers of digits which are* 1, will be zero. To see why, imagine that you move to the place where the last 1 occurs. Then consider the determination of the infinite tail sequence starting from that point. It follows that the property of having a finite number of ones is a *probability zero property* of infinite bit sequences.

Now let \( r \) be an arbitrary infinite bit sequence. We divide the bits of \( r \) into disjoint *groups* starting from the beginning, in such a way that the \( n \)th group consists of \( n \) bits. Below, the bits of every second group of \( r \) is written in boldface:

\[
r = 1010111000111101001011010010...\]

We call the natural number \( n \) a *winner with respect to \( r \) if \( n = 0 \) or the \( n \)th group of bits from \( r \) consists of 1’s only. In the example above, \( n = 1 \) and \( n = 5 \) are winners along with \( n = 0 \).

Consider the ordering

\[ E = \{ E_1, E_2, E_3, \ldots \} \]

induced on \( E \) by the ordering on \( E^+ \). Define the infinite binary sequence

\[ \tau \]

by letting group \( n \) of bits in \( \tau \) be the \( n \) first bits of the length \( k_n \) contribution sequence of the \( n \)th number \( E_n \in E \), where \( k_n \) is the place of \( E_n \) in the ordering of \( E^+ \). In other words,

\[ \text{ORD}_{E^+}(E_n, k_n) \]

is true in \( N \). Thus the first bit of \( \tau \) is the single first bit of the length \( k_1 \) contribution sequence of the first formula in \( E \), the next two bits are the first two bits in the length \( k_2 \) contribution sequence of the second formula in \( E \), and so on. By standard theory it follows that we may choose, or explicitly construct if we want, a sentence

\[ \text{INFW} \]

of \( L \) such that \( \text{INFW} \) is true in \( N \) iff there are an infinite number of winners with respect to \( \tau \).
5. Main result

We are now ready for the main technical result of this paper. Roughly speaking, this theorem states the following: There is a sentence \( D \) in \( L \) with the property that if an extension \( T \) of PA has \( N \) as a model part and decides \( D \), then either there is a natural number \( n \) such that \( T \) proves the existence of a zero in the length \( n \) contribution sequence based on its own NIN, or else \( T \) proves that there are infinitely many winners with respect to the sequence \( \tau \).

**Theorem.** There is a sentence \( D \) in \( L \) with the following property: If \( t \in E \) has \( N \) as a model part and \( t \) decides \( \lceil D \rceil \), then either there is a natural number \( n \) such that \( t \) proves \( \lceil \forall s (\text{nin}(t, s) \rightarrow \exists z \text{Contr}(s, n, z, 0)) \rceil \), or else \( t \) proves \( \lceil \text{INFW} \rceil \). The property of having an infinite number of winners is a zero probability property of infinite bit sequences.

**Proof.** Assume that \( t \in E \) decides \( \lceil D \rceil \). We are going to consider two cases, namely INFW true in \( N \) and INFW false in \( N \).

Assume first that INFW is true in \( N \). Let \( D = \text{INFW} \). Then since \( t \) has \( N \) as a model part, \( t \) cannot prove \( \lceil \neg D \rceil \). Since \( t \) decides \( \lceil D \rceil \), it follows that \( t \vdash \lceil D \rceil \).

Assume now that INFW is false in \( N \). We may choose (or explicitly construct) an \( L \)-formula \( \text{Winner}^+(x) \) with one free variable \( x \) such that for all natural numbers \( n \), the formula \( \text{Winner}^+(n) \) is true in \( N \) iff the length \( n \) contribution sequence of \( E_n^+ \) consists of bits 1 only, or \( n = 0 \). Since INFW is false in \( N \), there are only a finite number of bit groups in \( \tau \) consisting of bits 1 only. Since elements of \( E^+ \) which are not in \( E \) have contribution sequences consisting of bits 0 only, it follows that there are only a finite number of natural numbers \( n \) such that \( \text{Winner}^+(n) \) is true in \( N \). In other words, there exists a maximal natural number \( m \) with this property. Let \( D \) be the \( L(PA) \) sentence

\[
\text{Winner}^+(m) \land \neg \exists x (m < x \land \text{Winner}^+(x)).
\]

Thus \( D \) states that \( m \) is the “maximal winner”. By construction, \( D \) is true in \( N \). Note first that since \( D \) is true in \( N \), and \( N \) is a model part of \( t \), it is impossible that \( t \) proves \( \lceil \neg D \rceil \). Assume that \( t \) proves \( \lceil D \rceil \). Let \( t = \lceil A(x, y) \rceil \). Now by standard results there will be arbitrary large formulas \( B(x, y) \) with Gödel numbers in \( E^+ \) which are provably equivalent to \( A(x, y) \) in PA, and hence also in \( t \). Therefore, since PA is able to decide all numerical instances of the formula \( \text{ORD}^+(w, x) \), there is a formula \( B(x, y) \) with Gödel number in \( E^+ \) and a natural number \( n > m \) such that \( t \) proves both

\[
\lceil A(x, y) \leftrightarrow B(x, y) \rceil
\]

and

\[
\lceil \text{ORD}^+(\lceil B(x, y) \rceil, n) \rceil.
\]

Since \( t \) proves \( \lceil D \rceil \) and \( n > m \), it follows that \( t \) proves

\[
\lceil \forall s (\text{nin}(\lceil B(x, y) \rceil, s) \rightarrow \exists z \text{Contr}(s, n, z, 0)) \rceil.
\]
But since $t$ proves $[A(x, y) \leftrightarrow B(x, y)]$, it also proves that $[A(x, t)]$ and $[B(x, y)]$ share the same NIN. So $t$ proves

$[\forall s(\text{nin}(t, s) \leftrightarrow \text{nin}([B(x, y)], s))],$

it follows that $t$ also proves

$[\forall s(\text{nin}(t, s) \rightarrow \exists z \text{Contr}(s, n, z, 0))].$

Thus we have reached the conclusion of the theorem for the case where INFW is false.

To complete the proof, we must verify that having an infinite number of bit groups with bits 1 only, is a zero probability property for an infinite bit sequence $r$ with bits chosen successively at random. The definition of bit groups was given in section 4. Note first that the probability that our randomly chosen sequence has no winners $n > 0$, is

$$p = \frac{1}{2} \cdot \frac{3}{4} \cdot \frac{7}{8} \cdot \frac{15}{16} \cdot \frac{31}{32} \cdots$$

(3)

This infinite product, in which the $n$th factor is

$$\frac{2^n - 1}{2^n},$$

can be shown to converge to a nonzero number. We have $p \approx 0.289$. In line with the terminology of section 4, may summarize this by saying that the property of having no winners $n > 0$ is a nonzero probability property of infinite bit sequences. Moreover, the probability

$$p[n] = \frac{2^{n+1} - 1}{2^{n+1}} \cdot \frac{2^{n+2} - 1}{2^{n+2}} \cdot \frac{2^{n+3} - 1}{2^{n+3}} \cdots$$

that there are no winning numbers $j$ greater than $n$, tends to 1 as $n \to \infty$. Now imagine that we conduct the following infinite sequence of probability experiments with a given infinite bit sequence, examining its bits from the beginning upwards. (1) We look for the first winning bit group, starting from the first bit group. (2) If (1) succeeds, we look for the next winning group following the one we already found. (3) If (2) succeeds, we look for the next winner past the one we found in (2). And so on. If the sequence has an infinite number of winning groups, then all of these experiment stages will succeed. But it follows from the fact that $p[n] \to 1$ as $n \to \infty$ that each of these stages have a probability of failing which may be bounded uniformly from below by a nonzero number. By the law of large numbers, using a sequence in which bits are chosen successively at random, sooner or later we must encounter an experiment stage which fails. Then the sequence will not have an infinite number of winners. It follows that the property of having an infinite number of winners is a zero probability property of infinite bit sequences. ■

6. The absolute undecidability argument

The famous essay Proofs and Refutations by Imre Lakatos (Lakatos [1976]) is commonly referred to in arguing for the general fallibility of mathematical arguments. See, e.g., chapter 4 of Ernest [1998]. The argument is partly historical. Certainly the mathematical proof standard, i.e. the conditions
which must be met by an argument for it to be accepted as a “proof” by mathematicians, has changed over time. So then how can we be certain that the proof standard used by mathematicians today, will not also be revised some time in the future? In a historical perspective, it seems reasonable that we have to accept the idea that even the rigorous modern proof standard may be rejected by future generations. Such an argument can be given for any proof standard, hence mathematics is fallible.

A commonly used argument against this goes roughly as follows. The modern mathematical proof standard can be linked to formalizability in formal (first order) theories, in which correctness of proofs may be checked by an algorithm which halts on all inputs. Although most modern mathematicians are not logicians, and do not know how to formalize their proofs in a suitable first order theory, they have developed an intuition for recognizing arguments which are actually formalizable. These arguments are precisely the arguments which merely establish logical consequences of the given axioms, that is, they are the arguments establishing sentences which will be true in any structure (“interpretation”) where the axioms are all true. Among mainstream mathematicians, being able to recognize such arguments is typically referred to as “knowing what a proof is”, and developing this intuition is a central goal of mathematical education at the university level. Mirroring the “absolute” nature of the underlying algorithmic checkability, in modern mathematics there is no controversy concerning whether or not any concrete argument is a valid proof according to the modern proof standard or not. Proofs which are refuted today, are refuted because someone discovers an error in them. This error always takes the form of a “gap”, that is, some portion of the proof which relies on something more than pure logical consequences. One speaks of checking a proof, which again reflects the underlying algorithmic mechanism. Hence the modern proof standard is qualitatively different from previous proof standards, and there is no reason why it should be “revised” in the future. Arguing that there is a difference between “real” mathematics and its representation (model) in first-order logic (see, e.g. chapter 7 of Davis and Hersch [1981]) does not help, since that ignores the fact that the real-world mathematics proof standard is linked to first-order logic by the “knowing what a proof is” intuition. Also, arguing that there are (formalizable) concepts of proofs which differs from the standard one, such as intuitionistic proofs (see, e.g. chapter 1 of Ernest [1998]), may come through as irrelevant: There is no controversy concerning what is a proof according to intuitionistic criteria, and what distinguishes this from the the standard proof concept. A mathematician may work intuitionistically one day, and standard the next. Also, intuitionistic (or “constructive”) mathematics may even be simulated by formal theories using standard logic.

Based on the previous paragraph one might be tempted to conclude that the discussion of proof standards in Proofs and Refutations is primarily of historical interest. However, as I aim to demonstrate in this section, there is a twist to this. Even though I accept that the present-day mathematical proof standard is clearly “absolute” in the sense described above, and hence that proofs meeting this standard cannot ever be “refuted” in any meaningful way, the example I am about to demonstrate shows me that there is a different problem with this standard: It is too restrictive. It does not cover all convincing mathematical arguments. Hence, just like its historical predecessors, it too does not give a “final answer” to what mathematical reasoning is all about.

The argument which I am about to present, concludes that the sentence $D$ mentioned in the theorem of section 5 is absolutely undecidable in the sense that we will never be able to infer any truth value for it in $N$. This argument forms my example of a mathematical “proof” which is convincing, yet not formalizable. Its existence demonstrates the incompleteness of the modern proof standard (defined through formalizability) when it comes to covering all valid mathematical
arguments.

Naturally, there is a lot more to say about the status of the undecidability argument which follows. It may be expanded, and its logical status and inner nature can be illuminated by linking it to different theoretical frameworks in philosophy, in particular, of course, the philosophy of science. While I certainly do not think one should ignore this, my priority here has been on keeping the argument short, and through that hopefully making the main point come clearly through.

Before we commence, there is an elementary logical point which must be clarified. To avoid confusion concerning the following undecidability argument, one must keep in mind the difference between direct and indirect reference to formulas. Tracing back definitions, one may find that we can almost explicitly construct the formula \( D \) defined in the proof of the theorem from section 5. Indeed, if we had known the truth value of INFW in \( N \), and in case it is false, the value \( m \) of the maximal natural number satisfying Winner\(^*\)(\( x \)), we would have known which formula \( D \) is. Then we could construct it explicitly. However, we do not know these things. Therefore, the fact that we know from the proof by indirect reference that \( D \) is actually true in \( N \) in both cases considered, does not in any way help us finding “reasonable” axioms for a theory in which we can decide \( D \). In particular, we cannot include \( D \) as an axiom! However, the concrete formula \( D \) is out there somewhere. We may in principle stumble onto it, or write it down if you will. Then we may try to prove or disprove it in some theory with reasonable axioms, and it is the impossibility of succeeding in this I am going to argue for. An analogy: Knowing that the thief is by definition guilty (indirect reference), does not help you decide whether or not a person you meet on the street is guilty. I will argue that the formula \( D \) corresponds to a thief we can never catch. To put it yet another way: The concrete formula \( D \) is true, but we do not know that. I argue that we will never know.

The first step in the undecidability argument is to note that provability in any given mathematical theory \( T \) for which there exists a Gödel numbering of the language \( L(T) \) of \( T \) and a proof-checking algorithm, will be represented by a \( \Delta_0 \) formula \( A(x, y) \) in \( L \), where \( A(n, m) \) is true in \( N \) iff \( n \) is the Gödel number of a formula in \( L(T) \) and \( m \) is the Gödel number of a proof of \( n \) in \( T \). In other words, any such theory \( T \) will correspond to some number in \( E^+ \). Note that this setup is very general, we are not restricting ourselves to first-order syntax or anything like that. Roughly, we may say that we cover all mathematical theories which are formalizable and axiomatized. Note that theories \( T \) whose languages are not \( L(PA) \) may be dressed up as extensions of PA in our sense by adding axioms which can be used for translating theorems of \( T \) which are about natural numbers into equivalent formulas of \( L(PA) \). For instance, this may be done with ZF and other theories in \( L(ZF) \).

The second step is to consider the conditions defining the set \( E \subseteq E^+ \). I will argue that these conditions will be met by all the numbers in \( E^+ \) representing theories with “reasonable” axioms. Given \( t \in E^+ \), let \( t = \lceil A(x, y) \rceil \) where \( A(x, y) \) represents proofs in a theory \( T \). Translating the conditions on \( t \) required for \( t \in E \), we obtain the following two requirements:

(i) \( T \) does not prove \( \neg \text{Con}_T \), and if \( B \) is an \( L \)-formula such that \( \text{PA} \vdash B \), then \( T \vdash B \).

(ii) There is a sentence \( B \) in \( L(PA) \) such that (1) \( T \) does not prove \( B \), and (2) \( T \) does not prove \( \text{Con}_T \rightarrow \neg(T \vdash B) \).

Here (i) is clearly unproblematic: If \( T \) is a theory with reasonable axioms, then it cannot prove its own inconsistency. Moreover, we lose no generality in assuming that \( T \) is at least as strong as PA. This latter requirement justifies, of course, calling the numbers in \( E \) extensions of PA.
To see that (ii) must also hold if $T$ has “reasonable” axioms, assume that such a $T$ satisfies (i) but not (ii). Let $T' = T + \text{Con}_T$. Then $T'$ solves the general halting problem for Turing machines. To see this, note that since $T$ satisfies (i), $T \vdash B$ implies $T \vdash (T \vdash B)$. Given a Turing machine along with input, let $B$ be a formula expressing that it halts. Now again since $T$ satisfies (i), the machine halts iff $T \vdash B$ is true in $N$. But since (ii) is not satisfied, a search in $T'$ for a proof of either $T \vdash B$ or $\neg(T \vdash B)$ will always halt. This decides the halting problem.

Moving on, the third step in the undecidability argument is to consider the unique number $n \in N$ making $\text{nin}(m, n)$ true in $N$ for our given $m = [A(x, y)]$ in $E$ representing the theory $T$. I will call this $n$ the naturally inaccessible number (NIN) of the theory $T$. This number $n$ is the Gödel number of the smallest $L(\text{PA})$-formula $B$ such that (i) $T$ does not prove $B$, and (ii) $T$ does not prove $\text{Con}_T \rightarrow \neg(T \vdash B)$.

Concerning this, note the following little technicality: If there is a formula $B$ such that $T$ proves both

$$\text{Con}_T \rightarrow \neg(T \vdash [\text{Con}_T \rightarrow \neg(T \vdash B)]) \tag{4}$$

and

$$\text{Con}_T \rightarrow \neg(T \vdash B), \tag{5}$$

then from provability of (5) it follows that $T$ also proves

$$T \vdash [\text{Con}_T \rightarrow \neg(T \vdash B)].$$

Combining this with provability of (4), it follows that $T$ proves

$$\neg\text{Con}_T.$$

This implies that $T$ is not represented by a number in $E$. Hence, a (possibly non-minimal) candidate for the NIN of $T$ in the case where $T$ is represented in $E$, cannot be produced in $T$. Thus the NIN of $T$ is, in a sense, really inaccessible to $T$.

The final step in the undecidability argument is to assume that $T$ is a theory represented by $t \in E$, and that $t$ decides $D$. Then it follows from the theorem of section 5 that either (i) $T$ proves the existence of infinitely many winners with respect to $\tau$, or else (ii) there is a natural number $n$ such that $T$ proves the existence of a zero in the length $n$ contribution sequence based on its own NIN, a sequence which I will also refer to as the length $n$ contribution sequence of $T$. Let us call a theory reachable if it has axioms which are “reasonable”. I will argue that both of these situations are “absurd” if $T$ is reachable.

Consider (ii) first. Note that this is something entirely different from just proving the existence of a natural number $n$ such that the length $n$ contribution sequence of $T$ contains a zero. In contrast, the theory $T$ proves the existence of a zero in the contribution sequence of length $n$ with the number $n$ specified through a numerical term $n$.

But as pointed out above, our reachable theory $T$, even when assuming its own consistence, cannot even derive a concrete candidate for its own NIN. Now the length $n$ contribution sequence of $T$ is defined on basis of its NIN by performing some arithmetic operations on it and then picking out bits at selected, scattered positions close to the middle of the resulting number. The idea that
our reachable theory $T$ should be able to prove any relevant information concerning these bits, given its lack of knowledge about the NIN itself, may indeed be labeled “absurd”. Remember that for a random bit sequence of $n$ bits, the probability that the sequence has bits 1 only, is nonzero. Hence, some nontrivial information concerning the $n$ bits in the contribution sequence is needed to eliminate this possibility. It seems extremely unnatural to assume, even disregarding time/space limitations, that it is in principle possible to find axioms for a reachable theory achieving such a thing concerning its own NIN. However, if we take the position that $D$ is not undecidable by all the theories represented in $E$, then we are forced into making this absurd assumption. This is the decisive point.

Assume now instead that condition (i) holds, thus $T$ proves that there are an infinite number of winners with respect to $\tau$. Note first that $D$ expresses a zero probability property of infinite bit sequences. Also note that in $\tau$, contributions from elements $t \in E^+$ which do not satisfy the conditions we have imposed on $E$, are removed. The elements we are left with, all have nonzero NINs based on the Gödel numbers of minimal formulas not derivable in the “theories”. The contribution sequences thus are formed with nontrivial seeds, resembling pseudo random number generator designs. Taken into account the way in which $\tau$ is built from contribution sequences of the elements in $E$, using their NINs, the idea that one, concrete reachable theory $T$ should be able to prove a probability zero property such as this, across the full spectrum of the NINs of all elements in $E$, is again an absurd assumption.

All in all, this shows that unless we accept that $D$ cannot be decided by any theory with “reasonable” axioms, we are forced into making one of the two “absurd” assumptions we have now encountered. We cannot escape the problem. This completes the argument for the conclusion that there is no reachable theory deciding $D$, a conclusion which means that we will never be able to infer the truth value of $D$ in $N$. Our absolute undecidability argument for $D$ is complete.

Logically, the structure of the preceding undecidability argument is similar to classical reduction ad absurdum. However, since we are dealing with underdetermination instead of over-determination, the absurd results obtained are not contradictions. Instead, they describe absurd coincidences, or absurd situations. Without a formalizable definition of a “reasonable axiom”, we have no chance of formalizing the argument. But still, the argument which has been given for the impossibility of the situations in question, based on the definitions made, to me is as convincing as a proof establishing a logical consequence, a proof which then may be formalized and checked by a computer. The arguments are just different in nature. The beauty of the situation is that arguments for this kind of underdetermination are naturally not formalizable. The existence of convincing arguments of this kind shows that, contrary to what has been the mainstream view for more than a century, mathematics is not exclusively about about finding reasonable axioms and establishing logical consequences of them.

We close this section with a corollary which is a mere reformulation, but which represents a different way of viewing the undecidability argument.

**Corollary.** Let $E_0 \subseteq E$ be the set of all $t \in E$ which has $N$ as a model part, and which is such that (i) $t$ does not prove $\lceil \text{INFW} \rceil$, and (ii) there is no natural number $n$ such that $t$ proves the existence of a zero in the length $n$ contribution sequence based on its own NIN. Then there is a sentence $D$ of $L(\text{PA})$ such that no $t \in E_0$ proves or disproves $\lceil D \rceil$. 

14
Proof. This follows directly from the theorem in section 5.

The important point here is of course that the conditions defining $E_0$ are very weak. Their degree of weakness is exactly what has been discussed in this section. An interesting problem is to find even weaker conditions on $E_0$ which will also make the conclusion of the corollary go through. Of course, some slight weakenings are easy to see. For instance, one may replace “no natural number $n$” in (ii) with “only a finite number of natural numbers $n$”. Also, $\lceil \text{INFW} \rceil$ may be replaced by a stricter condition of the same “kind”. However, these weakenings are not very deep, and I presently do not know about any really substantial weakenings. A thing I find remarkable in this context, is the underlying role played by the convergent infinite product (3) in the conditions on $E_0$. This is an infinite product of numbers in $(0, 1)$ converging to a number $\alpha$ in $(0, 1)$. Such products have a kind of open-ended interpretation in terms of probability, an interpretation which in a way does not work. If we imagine an infinite string of experiments in which the probability of success in the $n$th experiment is equal to the $n$th factor of the product, then $\alpha$ in a way would represent the probability that none of the infinite number of experiments fail. So in a way, we have an infinite process with two different nonzero probability outcomes. In a way, this effect is deeply related to the construction of the formula $D$. I presently does not know about any method of constructing a similar sentence $D$, or, if you will, similar conditions for the set $E_0$ in the corollary, without involving, directly or indirectly, such a convergent infinite product.

Returning to the discourse on the existence of convincing, yet unformalizable mathematical arguments, the corollary illustrates the possibility of considering the unformalizable part of our undecidability argument simply as an informal motivation for the choice of new axioms. One could introduce a concept of reachable theory by means of axioms implying properties (i) and (ii) in the corollary, and one could then reformulate the corollary as an “absolutely undecidability theorem”. However, I feel personally that while this approach is certainly useful for systemizing things, a one-sided emphasis on it is misleading. When the theory is recast this way, (parts of) the absolute undecidability argument given in this section would comprise an example of a mathematically speaking highly nontrivial piece of motivation for the choice of axioms. Returning to the historical arguments commonly based on Lakatos (Lakatos [1976], Ernest [1998]), I certainly think it would be artificial if the mathematical reasoning involved in motivating things this way, was not recognized as real mathematics. So we arrive at the beforementioned conclusion that when taken alone, the present-day modern mathematical proof standard, going back to Cauchy approximately, is too restrictive for containing all forms of convincing mathematical reasoning. This shows that the “fallibilist” proof standard arguments based on Proofs and Refutations are, after all, not only of historical interest.

7. The foundations of mathematics and its link to reality

If there are sentences of $L = L(\text{PA})$ whose truth value in the standard structure $N$ can never be found, even in principle, then assuming that these sentences still have a (secret, non-constructible) truth value in $N$, to me is not natural. Thus to me, the example of $D$ shows that completeness of the truth concept in $N$ is in fact an unnatural assumption to make. Such completeness has traditionally been believed in by most mathematicians, but of course such beliefs are largely based on mathematical consequences. To me, the consequences described here change the picture. The example $D$ shows that assuming this kind of completeness leads to the existence of what can be
considered “absurd” results. So if we do not want to believe in such “absurd” results, then we cannot believe in the existence of a complete concept of truth for closed $L(PA)$ formulas in $N$ either.

If we give up the idea of working with a complete notion of truth in $N$ for $L(PA)$ formulas, then the entire idea about (complete) concepts of truth in infinite structures for formal languages will probably have to be abandoned. Thus we deviate from the general definition of truth deriving from Tarski [1956]. Looking in the direction of intuitionism and related developments (Brouwer [1981]) in a way does not help, since the underdetermination effect we experience exists even though the law of the excluded middle is allowed in the logic we use inside the system, that is, in our theories. However, if we jump out of the system and look at the concept of truth, it would appear that we are led to a kind of constructivist logic. The sentence $D$ from the theorem gives us example in which although $D \lor \neg D$ is trivially derivable and hence must be considered “true” in a constructive sense, it appears as though neither $D$ nor $\neg D$ has a meaningful (constructible) truth value. It is maybe possible to link this to the results on assertibility found in Weaver (2015).

The consequences of the above are not as far-reaching as one might first think. There are many ways out. For example, one may adopt the simple view that each consistent formal theory $T$ defines one mathematical object (model) $M_T$, which one could call the generic model of the theory $T$. One could justify this by adopting a proof of Gödel’s completeness theorem in the version “Every consistent theory has a model”. A mathematical object then by definition would be the generic model $M_T$ of some consistent formal theory $T$. In $M_T$, truth is defined as provability in $T$. Formulas which are undecidable in $T$ will have an undefined truth value in $M_T$. For instance, if $T$ is a theory describing the general concept of a group, then the generic model $M_T$ of $T$ could be called the generic group. More generally, by a model of $T$ we would mean some mathematical object where all axioms in $T$ are true. (Remember now that each of these objects will have concepts of truth defined by provability in the theories for which they are the generic model.) In this example, it would be natural to call the models of $T$ groups. As an example, consider the extension $T'$ of $T$ describing the group $S_5$. Then the concept of truth in $S_5$ relative to the language in question is defined by provability in $T'$. Since all axioms of $T$ hold in $S_5$ according to this definition of truth in $S_5$, the object $S_5$ is a group. Thus we may still speak of “the group $S_5$”. In other words, mainstream mathematics terminology is left more or less unchanged.

8. Why is this important?

But if more or less nothing changes, then one may rightfully ask why the paradigm shift I argue for, is necessary. Why can’t we just continue using our romantic idea of a complete concept of truth concerning natural numbers, as long as this does not lead to inconsistency?

First of all, once one realizes that this idea has absurd consequences, it is not so romantic any more. Secondly, maintaining a view of mathematical objects which is (arguably) unnatural, may have the effect that mathematical thinking is led away from ideas which may reflect important fundamental properties of the physical world, real life and indeed mathematics itself. History is filled with examples of surprising links between mathematics and other subjects. Readers acquainted with physics may have noticed the obvious analogy between quantum theory and the view of truth in mathematical structures suggested above: Mathematical objects only exist as defined by our knowledge of them, and while $A \lor \neg A$ is always true, it can happen that neither $A$ nor $\neg A$ is true.
This corresponds to the principle of superposition in quantum physics. Certainly it is possible to imagine also that the underdetermination effect described by the example of $D$ may have physical counterparts. It should be enough to mention the problem of free will in human brains. How can the manoeuvring space for this exist if the world is governed by physical laws which are finitely describable at the micro level? The answer may turn out to be that some physical systems, despite being of finite size, have an infinite complexity buried in them which makes their actions underdetermined by the physical laws in question. Such a system would then form an analogy to the formula $D$ in this paper.

REFERENCES


