Introduction

The life of the mathematician is governed by the following two simple principles: for any mathematical assertion A, if it has a proof it is true, and conversely, if A is true – that is, if A is to be granted the status of a *theorem* – A must have a proof.

Into this neat correspondence between mathematical truth and mathematical proof, indeed into this soundest and most venerable of all the human sciences, namely mathematics, Gödel introduced a small, filament-like crack – albeit with a number of essential caveats. And whether mathematicians will ever have to face the consequences of Gödel's theorems, or whether a *cordon sanitaire* exists around mathematics, immunizing its core areas from the effects of incompleteness, as some have conjectured,[1] the fact is that mathematics had changed forever after 1931 – for those who cared to ponder the matter.

The Incompleteness Theorems: what are they about? What do they say about the mathematician's everyday concept of truth – not to mention everyone else's? Do the Incompleteness Theorems have even broader implications, for example for the computational nature of the mind, or for quantum physics, or, as some have suggested, for the organization of society and the logical consistency of self-government? Concerning the proof, which is based on the Liar Paradox, is it a "parlor trick," as Gödel suggested to Kreisel on a walk one day?[2] Or is it a masterpiece of logical methodology?

The logician has questions of their own: are the Incompleteness Theorems fundamentally about self-reference, about the existence of fixed points for formulae in, say, the language of arithmetic? Do they rely essentially on diagonalization and paradoxes? Or are there diagonal/self-reference–free proofs of them? Can semantic notions such as truth and definability be eliminated from the statement of the theorems and their proofs and, if so, are the syntactic versions purely syntactic? How tied to the language of arithmetic, or indeed to any particular language, are they? To which formal systems do they apply?

In this Element we weigh in on the logician's sharper questions, setting aside the broader questions raised by political scientists, theologians, and others.[3] In spite of Gödel's early worries about their generality, the Incompleteness Theorems have turned out to be incredibly plastic, appearing in as many different (dis)guises and with as many different proofs as a Greek god in pursuit of *amour*: syntactic, semantic, abstract or logic-free, liar-free, language-free, "honest," that is, such that coding and representability are presented in

[1] See Macintyre's [92].
[2] See Kreisel's Royal Society obituary of Gödel [80].
[3] An exception is the Lucas–Penrose debate, which we (briefly) take up in Section 8.

complete detail; "mathematical," treelike, arithmetic, nonarithmetic, proof by recursively inseparable sets, proof by undecidability of the halting problem, and so forth. And although both theorems are slightly *unstable*, in the sense that with a little logical trickery in the form of so-called deviance, for example, deviant provability predicates, or deviant notation systems, they can be made to fail,[4] they are at the same time remarkably robust.

Of particular interest to us is the classification of the various proofs of the Incompleteness Theorems into the categories of syntactic and semantic – categories, as we have argued elsewhere, that can be somewhat changeable.[5] Accordingly we here explore the phenomenology of the syntax/semantics distinction, as seen through the prism of the Incompleteness Theorems. About the semantic concept of truth in particular, it is striking that in spite of his embrace of truth as a primitive notion in the 1940s, Gödel goes to great lengths to banish the concept of truth from the proof of the Incompleteness Theorems – in deference, as he would later admit, to the anti-metaphysical tenor of the times. We will also emphasize the role the theorems played in the transition in logic, and especially in set theory, from a type-theoretic framework to a first-order one. Gödel's *1931*[6] very evidently tracks this change.

If one looks back *comprehensively* on the logic of the 1920s, one can detect slight indications coming from here and there that something on the order of Gödel's *1931* was coming, especially from E. Post, and, for the Second Incompleteness Theorem, from Kuratowski.[7] It must be said, however, that many logicians, including Hilbert, hoped and indeed expected the opposite of the results of *1931*. All of this would be swept aside; on the side of premonition, the technical machinery Gödel invented for the proof went far beyond what had been done to date; and on the side of those expecting to achieve the goals of the Hilbert Program, the proof left no doubt that the desiderata of that Program, as stated, could not be carried out—shaking that program to its foundations, as Kleene would write, without quite demolishing it.[8]

It is hard to imagine the undeveloped state of logic prior to 1931 – or nearly so, considering Tarski's 1926–8 Warsaw seminar on the elimination of quantifiers, if not the Polish School in logic altogether; or considering Gödel's 1929 doctoral thesis. Computability theory, which is intimately bound up with the two Incompleteness Theorems, was in its infancy—in fact Gödel's *1931* was a

[4] See Section 4.1.
[5] See [72].
[6] Throughout this Element we follow the numeration of Gödel's papers as given in the *Collected Works*, vols. I–V. So "*1931*" refers to Gödel's 1931 paper, while "1931" refers to the year 1931.
[7] See Section 5.1.
[8] [42], p. 127.

central stimulus for the development of that field. The notion of a formal semantics or even the logical notion of a model did not exist at the time, although semantic notions were relied on *sotto voce*, more or less, to wit: the structure of the natural numbers as it is put to use in Gödel's thesis, and, more overtly, albeit "piecewise" in Theorems V, IX, and X of Gödel's *1931*. The notion of truth was famously and vociferously campaigned against by, for example, the Vienna Circle, even so that the everyday mathematician could still entertain the idea that in mathematics, truth simply meant the same thing as proof—and indeed for today's intuitionist this remains so. Logicians conflated first- and second- (or higher) order logic, which Gödel's 1929 Completeness Theorem definitively separates. Finally, the concept of "finite computation" was not well understood. Gödel and Herbrand in particular expressed doubts that an adequate definition of the concept could be given, doubts which were only put to rest in Gödel's mind in 1936 with the advent of the Turing Machine.

On the level of syntax and semantics, one would be hard-pressed to find the distinction laid out in, for example, Russell and Whitehead's *Principia*, which is essentially an interpreted system, whereas Gödel's *1931* draws a sharp distinction between the two.[9] Here is Gödel patiently (or perhaps impatiently) explaining the distinction between name and referent in a 1931 letter to Zermelo, himself a figure of colossal importance for logic and set theory, in response to Zermelo's questioning the proof of the Incompleteness Theorems:

> Namely, one can not set
>
> $$n \in K^* = \overline{[R(n); n]},$$
>
> because the symbol complex $[R(n); n]$ has no meaning. A negation stroke, after all, only has meaning with reference to a symbol complex that expresses an assertion (with reference to the number 5, say, a negation stroke is meaningless). But the symbol complex "$[R(n); n]$" does *not* express an *assertion*. "$[R(n); n]$" means about the same as the following [English] words: "that formula of *Principia Mathematica* which results from the n-th class sign by substitution of the number n for the variable." "[R(n);n]" is not itself that formula . . . those words, however, obviously express no assertion, but are rather the unique characterization of a formula (that is, of a spatial figure), just as, say, the words "the first formula of that book" express no assertion, even if perhaps the formula that is characterized by those words does express an

[9] Grattan-Guinness in [52], p. 296 describes the situation thus.

> Despite the attention given to the distinction between theory and metatheory in the 1920s with Hilbert's revival of his proof theory, logicians still tended to conflate symbol and referent; indeed, Professor J. Barkley Rosser once told me in reminiscence that it was only with Gödel's theorem that logicians realised how careful they needed to be in this matter.

> assertion. For each particular number n, "$[R(n); n]$" thus is a name (a unique description) for a particular formula (i.e., a spatial figure), and a negation stroke over it therefore has just as little meaning as [it would], say, over the formula "$5 + n$," which, for every number n, is a name for a particular natural number. The whole difficulty obviously is due to the fact that in meta-mathematics there are, besides the symbols for numbers, functions, etc., also symbols for formulas, and that one must clearly distinguish a symbol that denotes a formula from that formula itself.[10]

Does mathematics remember what happened to it in 1931? Almost 100 years after their publication, the Incompleteness Theorems appear to have had a minor impact on the life of the working mathematician. What the theorems have to say about the undecidability of certain Diophantine equations, for example, has not really come to the surface outside of logic, if it ever will.

Mathematics goes on in spite of those theorems – a testament, perhaps, to its great structural stability.

[10] [47], pp. 425–427.

THE FIRST INCOMPLETENESS THEOREM

1 The First Version of the Proof

Gödel's proof of the First Incompleteness Theorem asks the viewer to shift their perspective back and forth, from semantics to syntax and back again to semantics, and back again to syntax. It is a hall of mirrors such as had never before been seen in mathematics – though diagonal arguments, of which Gödel's is a supreme example, had been around at least since the time of Cantor's proof of the uncountability of the real numbers (if not earlier).

Gödel's original, informal, and unpublished (in 1931) proof of the First Incompleteness Theorem was semantic in flavor, based as it was on the undefinability of truth versus the definability of the concept of provability with respect to the arithmetic system S in which he worked, together with the soundness of the system S. The observation is simply that the set of all S-provable sentences is a subset of the set of all sentences in the language of S true in the natural numbers – in fact the former is a definable subset of the latter. But the set of all sentences in the language of S true in the natural numbers is not definable, on pain of paradox. Thus the set of all S-provable sentences is a *proper* subset of the set of all sentences in the language of S true in the natural numbers.[11] From Gödel's 1964 letter to van Heijenoort:

> Perhaps you were puzzled by the fact that I once said an attempted relative consistency proof for analysis led to the proof of the existence of undecidable propositions and another time that the heuristic principle and the first version of the proof were those given in Sect. 7 of my 1934 Princeton lectures. But it was precisely the relative consistency proof which made it necessary to formalize either "truth" or "provability" and thereby forced a comparison of the two in this respect. By an enumeration of symbols, sentences and proofs within the given system, I quickly discovered that the concept of arithmetic truth cannot be defined in arithmetic. If it were possible to define truth in the system itself, we would have something like the liar paradox, showing the system to be inconsistent …Note that this argument can be formalized to show the existence of undecidable propositions without giving any individual instances. (If there were no undecidable propositions, all (and only) true propositions would be provable within the system. But then we would have a contradiction.)
>
> In contrast to truth, provability in a given formal system is an explicit combinatorial property of certain sentences of the system, which is formally specifiable by suitable elementary means.[12]

Gödel's conclusion that truth cannot be defined internally in S depends on *arithmetization*, namely an injective mapping from finite strings of symbols

[11] See Theorem 1.0.1 for the exact proof of this remark.
[12] See [47], p. 313.

into the natural numbers $\mathbb{N}$,[13] together with the *Fixed Point Theorem*. We explain these two concepts in Sections 2.1.2 and 2.1.5 respectively. Assuming these two concepts are in place, and letting $\ulcorner\phi\urcorner$ stand for the so-called "Gödel-number" of ϕ (see below), the undefinability of truth follows:

Theorem 1.0.1 *Let $\mathcal{T}$ be the set $\{\ulcorner\phi\urcorner \mid \mathbb{N} \models \phi\}$.*[14] *Then there is no formula $\theta(x)$ in the language of arithmetic such that for all ϕ:*

$\mathbb{N} \models \theta(\ulcorner\phi\urcorner)$ *if and only if* $\mathbb{N} \models \phi$.

Proof. Assume to the contrary that there is such a $\theta(x)$. Let ϕ be a fixed point of $\neg\theta(x)$.[15] That is,

$\mathbb{N} \models \neg\theta(\ulcorner\phi\urcorner)$ if and only if $\mathbb{N} \models \phi$.

If $\mathbb{N} \models \phi$ then $\mathbb{N} \models \theta(\ulcorner\phi\urcorner)$, a contradiction, by the definition of θ. If $\mathbb{N} \models \neg\phi$ then $\mathbb{N} \models \theta(\ulcorner\phi\urcorner)$, by the definition of ϕ. This is also a contradiction, as $\mathbb{N} \models \neg\phi$ implies $\mathbb{N} \models \neg\theta(\ulcorner\phi\urcorner)$, by the choice of θ. □

With the above theorem in place, Gödel could now derive the initial version of the First Incompleteness Theorem. Thus if $\mathcal{P}$ denotes the set of all Gödel-numbers of statements provable in S (under some suitable coding of those sentences fixed in advance), and $\mathcal{T}$ is again the set $\{\ulcorner\phi\urcorner \mid \mathbb{N} \models \phi\}$, where $\ulcorner\phi\urcorner$ denotes, as above, the Gödel-number of ϕ, then:

Theorem 1.0.2 *$\mathcal{P}$ is a proper subset of $\mathcal{T}$.*

Proof. By soundness, if $\ulcorner\phi\urcorner \in \mathcal{P}$, i.e., ϕ is provable, then ϕ is true. Thus $\ulcorner\phi\urcorner$ is an element of $\mathcal{T}$, i.e., $\mathcal{P}$ is a subset of $\mathcal{T}$. $\mathcal{P}$ is definable in S by a formula in the language of S, recalling Gödel's remark above that "In contrast to truth, provability in a given formal system is an explicit combinatorial property of certain sentences of the system, which is formally specifiable by suitable elementary means." If $\mathcal{P}$ were identical to $\mathcal{T}$, then $\mathcal{T}$ would be definable in S by that same formula. But $\mathcal{T}$ is not definable by *any* arithmetic formula, by the above Theorem 1.0.1. Thus $\mathcal{P} \subsetneq \mathcal{T}$. □

The theorem that truth is undefinable in the above sense is usually attributed to Tarski, who published the proof in 1933 and the German translation in 1936 [124]. Gödel clearly had Tarski's theorem in some form already in 1930. As for the question of priority related to the theorem on the undefinability of truth,

[13] In this Element the notation "$\mathbb{N}$" is used to denote both the set of natural numbers and the standard model of arithmetic. It will be clear from the context which is meant.
[14] The notation "$\mathbb{N} \models \phi$" means that ϕ is true in the standard model $\mathbb{N}$.
[15] For a proof of the Fixed Point Theorem, which Gödel clearly knew in 1930, see Section 2.1.5.

Gödel refers to Tarski in a footnote added at this point in the 1965 reprinting of his 1934 Princeton Lectures, albeit somewhat acidly.[16]

In these 1934 Princeton Lectures Gödel ponders self-reference, remarking that Russell and Whitehead's prohibition of *all* self-referential statements is "too drastic."[17] He cites the Fixed Point Theorem as a counterweight, as it provides a way, given "any metamathematical property *f* which can be expressed in the system, to construct a proposition that says of itself that it has this property."[18]

Gödel inserts the "first version of the proof," as he calls it, into section 7 of his 1934 Princeton Lectures, as mentioned in the above quote. Why didn't Gödel publish this first version, or even mention it, in his 1931 paper?[19] He later ascribed this to the "philosophical prejudices of the time":

> I have explained the heuristic principle for the construction of propositions undecidable in a given formal system in the lectures I gave in Princeton in 1934 …The occasion for comparing truth and demonstrability was an attempt to give a relative model-theoretic consistency proof of analysis in arithmetic. This leads almost by necessity to such a comparison.
> However in consequence of the philosophical prejudices of our times 1. nobody was looking for a relative consistency proof because i[t] was considered axiomatic that a consistency proof must be finitary in order to make sense. 2. a concept of objective mathematical truth as opposed to demonstrability was viewed with greatest suspicion and widely rejected as meaningless.[20]

Gödel often spoke in this vein, that is, about the philosophical or positivistic prejudices of the time, and indeed his published proof demonstrates the great care he took in order to accommodate such prejudices. Unlike his initial, informal proof, the published proof avoids to the degree possible semantic notions such as the notion of a model, of soundness, or of satisfaction in a

[16] "For a closer examination of this fact see A. Tarski's papers 1933a…" [42], p. 363. For more on the question of priority with regard to the Undefinability of Truth Theorem see Feferman's [22]. See also Woleński's [138]. Gödel also may have gone some way beyond Tarski's theorem in 1931 in the matter of model-theoretic semantics, based on the evidence of the mysterious footnote 48a of *1931*. On this point see Gödel's correspondence with Zermelo in the *Collected works* [47]. We discuss footnote 48a in Section 2.4.1.

[17] [42], p. 362.

[18] [43], p. 362.

[19] The heuristic or, in Gödel's words, "nonbinding" proof sketch of the First Incompleteness Theorem that Gödel gives at the beginning of *1931* resembles more the proof from a noncomputable set given in Section 2.5.1 than the original proof, though it does employ the concept of soundness. See Gödel's correspondence with Zermelo in [47] for Gödel's use of the term "nonbinding."

[20] Gödel's unsent letter to Balas, undated, [46], p. 10.

structure.[21] The caveat "to the degree possible" is meaningful; Gödel relies on a somewhat semantic condition of ω-consistency in his proof, an assumption that was later eliminated by Rosser [113], and the published syntactic proof has other semantic aspects, which we take note of as the particular issue arises.

Feferman explains Gödel's avoidance of the concept of truth in his *1931* thus, citing the above excerpt from Gödel's unsent letter to Balas:

> Here, in a crossed-out passage in an unsent reply to an unknown graduate student, I think we have reached the heart of the matter. Despite his deep convictions as to the objectivity of the concept of mathematical truth, Gödel feared that work assuming such a concept would be rejected by the foundational establishment, dominated as it was by Hilbert's ideas. Thus he sought to extract results from it which would make perfectly good sense even to those who eschewed all nonfinitary methods in mathematics ... Even more, once Gödel realized the generality of his incompleteness results it was natural that he should seek to attract attention by formulating them for the strong theories that had been very much in the public eye: theories of types such as PM and theories of sets such as ZF (Zermelo-Fraenkel). But if the concept of objective mathematical truth would be rejected in the case of arithmetic, should not one expect an even greater negative reaction to the case of theories of types or sets? All the more reason, then, not to have any result depend on it, and no need then to express one's convictions about it.[22]

Feferman's analysis notwithstanding, there is always the (rather remote) possibility that Gödel excised the concept of truth from the First Incompleteness Theorem on the basis of hesitations of his own that he may have held at the time. Seen in this light, the First Incompleteness Theorem may provide evidence for an anti-truth stance, however short-lived.[23]

---

[21] Gödel does invoke soundness in the informal sketch of the proof given at the beginning of *1931*. Also, Gödel's Theorem V relies on the concept of "piecewise" truth in the standard model. We discuss this point below in Section 2.3.1.

[22] [24], pp. 160–161.

[23] Evidence for this is somewhat sparse in the record, but it is there. See, for example, Gödel's remark in his 1933 Cambridge address:

> The result of our previous discussion is that our axioms, if interpreted as meaningful statements, necessarily presuppose a kind of Platonism, which cannot satisfy any critical mind and which does not even produce the conviction that they are consistent.

[45], pp. 49-50. For a fuller discussion of this matter see Davis's "What did Gödel believe and when did he believe it?" [13]. See also Parsons, Platonism and mathematical intuition in Kurt Gödel's thought [107]. We do not attend here to the subtleties regarding the nature of Gödel's Platonism, its relation to truth, and in particular to what Gödel himself may have meant by the term in 1933.

Putting the role of truth aside for the moment, Gödel's piercing clarity regarding the syntactic and semantic aspects of the proof can be seen in the fact that Gödel is always scrupulous to separate those elements of the proof that, in his words, "have nothing to do with the formal system P"[24] from the formal concepts. In the instance just quoted, Gödel is observing that the concept of a number-theoretic function being defined recursively from other number-theoretic functions is, simply, formalism-independent. This as it might be called "schematic" or modular approach made it easy for logicians to generalize the proof to a range of formal *and informal* settings subsequently, an enterprise that continues vigorously to this day.

2 Gödel's "Intuitionistically Acceptable" Second Proof of the First Incompleteness Theorem

2.1 Ingredients of the Proof

In our presentation of Gödel's 1931 proof we largely conform to Gödel's original presentation, though we adopt, as is usually done, a first-order setting rather than working in type theory. For a treatment of the proof including any details omitted here the reader is referred to Lev Beklemishev's brilliant survey [4], and to Kleene's introduction to [40], two of the deepest commentaries on the Incompleteness Theorems in the ocean of literature on them.

The formal system for which the Incompleteness Theorems were proved is a version of the simple (unramified) theory of types with the (second-order) Peano axioms adjoined. Using rather *PA* as the base theory,[25] the ingredients of the proof are as follows.

2.1.1 ω-consistency

We let $\overline{n}$ stand for the numeral term $s(s(\cdots s(\overline{0})\cdots))$, where s is the successor function symbol of the theory *PA*, the number of applications of the function s is n, and $\overline{0}$ denotes the constant term zero. We now state the definition of ω-consistency.

Definition 2.1.1 Let T be a theory extending *PA*.[26] We say that T is *ω-consistent* if it is not the case that there is a formula $\phi(x)$ in the language of T such that the following hold simultaneously:

[24] [40], p. 157.
[25] See the Glossary for the signature of *PA*, and a list of the Peano axioms.
[26] Alternatively, T can also be one of the *weaker* theories Q or R. See the Glossary for the definition of Q and R.

(i) $T \vdash \exists x \phi(x)$.
(ii) For all n, $T \vdash \neg\phi(\overline{n})$.

Note that ω-consistency is weaker than (the semantic condition of) soundness; at the same time it is stronger than (the syntactic condition of) mere consistency. Of course, ω-consistency implies consistency.

2.1.2 Arithmetization/Gödel-Numbering

Arithmetization (a.k.a. Gödel-numbering) is a one-to-one mapping of finite formal strings of the language of *PA* into the natural numbers. In Gödel's 1931 coding every element of the signature is assigned a distinct odd number; strings of symbols are then encoded with the help of the prime numbers via the mapping: $\langle n_1, n_2, \ldots, n_k \rangle \mapsto 2^{n_1} \cdot 3^{n_2} \cdots p_k^{n_k}$, where p_k is the k-th prime. In this way any formula or sentence ϕ corresponds to a number, its Gödel-number $\ulcorner \phi \urcorner$; and given any natural number, the formula it encodes (if the number is the Gödel-number of a formula) can be recovered from its Gödel-number because of prime factorization. Now it is easy to see that finite sequences of formulas can be encoded; notably, finite *proofs* can now be assigned a Gödel-number.

When it comes to the *definability* of coding, there is a subtlety here involving the β-function, so-called (see below). The above coding is not in any obvious way definable (unless we have the exponential function in the vocabulary) and while the (iterated) Cantor pairing function can be used to encode a finite sequence of any given length n, for each n the formula used to encode the sequence is different.[27] However with the β-function one has a single formula in plus and times that can encode sequences of any given length. Gödel uses the β-function later in the paper to show that the undecidable sentences constructed therein are arithmetic.

2.1.3 Primitive Recursion

Gödel defines the class of primitive recursive functions, as they are now known, or, as Gödel calls them, the "recursive" functions. In contrast to Gödel-numbering, which was a complete innovation at the time, primitive recursion was known.[28] The primitive recursive functions are defined as follows:

[27] See Section 2.3.1 for the definition and an application of the β-function.
[28] The phrase "primitive recursive" was coined by Rózsa Péter in 1928, the year Wilhelm Ackermann proved the existence of a recursive function that is not primitive recursive. See [3]. The primitive recursive functions appear as early as in Richard Dedekind's 1888 *What are numbers and what should they be?* [17].