Pázmány Papers Vol. 2, Nr. 1 (2024)
ISSN 3004-1279
https://doi.org/10.69706/PP.2024.2.1.3

Undecidability and the Reference of Formal Systems

János V. Barcsák[1]

Abstract

This paper attempts to provide an account of the reference of formal systems. I assume (on grounds that I cannot lay out fully) that formal systems can be considered to be referential, that is, capable of formulating truths in the correspondence sense, on two conditions: 1. that they are consistent and 2. that they contain true but unprovable formulas. The first of these conditions is self-evident; the second, by contrast, cannot be assumed without begging the question, without presupposing truth before accounting for its possibility. I argue, however, that Kurt Gödel's proof of the inevitability of undecidable formulas in any formal system provides a ground for assuming the existence of true but unprovable sentences without presupposing objective truth. For this, however, we need to develop a different sense of 'true' from what is usually assigned to the undecidable formula. Using insights from Jacques Derrida, I argue that we can legitimately conceptualize the truth of the undecidable formula as referring not to some objective reality but to the formal system itself.

Keywords

Kurt Gödel, Jacques Derrida, undecidability, correspondence truth, reference

1. Introduction: Derrida and Gödel

Unlike many of his contemporaries, Derrida rarely speaks about formal logic or mathematics. Several French philosophers of his generation – such as Lacan, Deleuze or Kristeva – are apparently attracted to mathematical analogies, and Badiou bases his whole theory of the subject, of the event, and of truth procedures on formal logical considerations. In spite of his conspicuous silence about logic and mathematics, however, Derrida makes a remarkable reference to Gödel's undecidable sentences when introducing his own notion of undecidability (Derrida 1981b, 230). This, as Christopher Norris

[1] Pázmány Péter Catholic University, barcsak.janos@btk.ppke.hu

remarks, is a telling invocation of Gödel's incompleteness theorem, since it occurs at a cardinal point, "notably in [Derrida's] treatment of Mallarmé's paradoxical reflections on language, logic, reference and truth" (Norris 2012, 34), that is, at a point where Derrida is engaging with the most foundational issues of deconstruction. Derrida's allusion to Gödel might, therefore, indicate an analogy between the formal-logical, metamathematical Gödelian argument about undecidable formulas of arithmetic and the fundamental strategies of deconstruction. This analogy has been explored in some detail in literature that attempts to create a link between deconstruction and analytic philosophy, most notably by Graham Priest, Paul Livingston, and Christopher Norris (Priest 2002; Norris 2012; Livingston 2012).

Although Priest does not compare Derrida to Gödel specifically, his comments on deconstruction can no doubt make such a comparison possible. In his 1994 paper titled "Derrida and Self-Reference" he likens deconstruction's emphasis on the inexpressibility and unnameability of its central terms (such as *différance*) to the early Wittgenstein's thoughts on ineffability (Priest 1994), and in his 1995 book *The Limits of Thought* he presents the fundamental strategy of deconstruction – essentially on the basis of the same analysis – as an instance of what he terms the Inclosure Schema. The Inclosure Schema is a set of conditions that results in the production of a specific kind of contradiction wherein a term or a member of some totality is both inexpressible in terms of the theory organizing that totality (Transcendence) and is nevertheless expressed or conveyed by that theory (Closure). Priest discovers this schema and the resulting contradictions in the work of a great number of thinkers throughout the history of philosophy, including of course Gödel, whose undecidable formula he presents as an inclosure contradiction, since its undecidability both transcends the theory of provability in terms of which it is formulated and acquires its sole formulation in terms of this theory (Priest 2002, 144).[2] Similarly, he sees Derrida's central (non-)concepts, *différance*, trace, supplementarity, *pharmakon*, *parergon*, etc. as manifesting the same kind of contradiction. These (non-)concepts are inexpressible in terms of the context in which they emerge, since they transcend the founding opposition organizing that context, and yet, precisely by this inexpressibility, the context in which they appear still succeeds in conveying a sense of what is inexpressible. What is more, this inexpressible can only be revealed inside this context, albeit only as that which transcends it. It is by this means that the deconstructive procedures organized around these (non-)concepts satisfy the two

[2] For Priest's in-depth discussion of Gödel's formula see his (Priest 2006, 39–50).

main conditions of Priest's Inclosure Schema: Transcendence and Closure (Priest 2002, 214–224).

Christopher Norris criticizes Priest's interpretation of Derrida for not taking into account how persistently the latter insists on consistency and on a classical bivalent logic (Norris 2006, 50; Norris 2012, 138; 148–149). However, he also recognizes and puts special emphasis on the analogy between deconstructive procedures and Gödel's incompleteness theorems. He points out that the aporetic outcome of "the various modes of deconstructive close-reading […] can best be understood by analogy with Gödel's incompleteness or undecidability theorem" (Norris 2012, 11), and insists that Derrida's invocation of Gödel's theorem "is not just a vaguely analogical or downright opportunist appeal to the presumed authority of mathematics and logic but a reference-point that precisely captures the movement – the logico-syntactic-semantic procedure – of Derrida's classic readings" (Norris 2012, 28). In spite of these general claims, however, Norris does not actually describe this analogy in detail. He makes a strong case for the relevance and applicability of formal logical considerations to deconstruction and vice versa, but the connection between Gödelian and Derridean undecidability remains merely implicit.

We are given a much more explicit treatment of this connection in Paul Livingston's book *The Politics of Logic*.[3] Similarly to Norris, Livingston starts out from the observation that "several of Derrida's key terms (for instance, *trace*, the 'undecidable,' and *différance*) and the textual *praxis* they embody can indeed usefully be understood as figuring the metalogical consequences of a thoroughgoing reflection on the implications of formalism" (Livingston 2012, 113). He then analyses Derrida's thoughts on mimesis (as expressed in his reading of Mallarmé in *The Double Session*) and on the term *différance* in close analogy with the status and function of Gödel's undecidable formula in the context of formal systems. He concludes that we can discover three fundamental similarities between Derrida's key terms and Gödel's undecidable formula:

> First, both depend on a kind of "self-referential" encoding whereby a system's total logic (the conditions for the possibility of its organizing distinctions) is formalized at a single point – the Gödel sentence or the "undecidable term" – which in turn makes it possible to inscribe an "undecidable." Second, both suggest a *generalization* of this result to show that any system of sufficient complexity will allow the inscription of undecidables […] [And third, both Gödel and Derrida's

[3] Cf. also his paper "Derrida and Formal Logic: fomalizing the undecidable" (Livingston 2010), which is the original of the chapter in *The Politics of Logic*.

> undecidable] always results from a semantical effect of *syntax* that cannot itself be excluded from any regular system of writing. (Livingston 2012, 121–122)

I will obviously not be able to represent the depth of Livingston's comparison here.[4] Suffice it to say that his detailed and carefully laid-out argument clearly establishes a deep connection between Gödel and Derrida's strategies. In this paper I will explain my own interpretation of this connection. I will present in outline an argument from a book that I am still working on; an argument in which I will attempt to show that Derrida's insightful treatment of undecidability can ground a new approach to the old problem of the reference of formal systems. This means that I will reverse the typical approach to the connection between Derrida and formal thought. Priest, Norris, and Livingston essentially use the analogy with formal logical considerations to provide a deeper understanding and justification of Derrida's arguments. Although both Norris and Livingston point out that analytic philosophy has much to learn from deconstruction, neither makes the case that deconstruction can have any bearing on formal logic. My starting point, on the other hand, is precisely this. I will argue that the development of formal logic has posed philosophical questions which can perhaps be handled in novel ways by implementing some of Derrida's insights.

In his seminal essay, "Différance" Derrida contends that "*différance* lends itself to a certain number of nonsynonymous substitutions, according to the necessity of the context" (Derrida 1981a, 12), and I will argue that Gödel's undecidable formula can be thought of as one such nonsynonymous substitution. The context in which this substitution occurs, moreover, is a particularly well-defined and lucid one: that of formal logical systems, which means that examining the status of Derridean undecidability in this precise context can also bring us closer to realizing Derrida's ambition expressed in his "Afterword: Toward an Ethic of Discussion": namely, to achieve "*the strictest possible determination* of the figures of play, of oscillation, of undecidability" (Derrida 1988, 145 (my italics)).

2. Formal Systems

Let us begin by familiarizing ourselves a little with the context: the precise and well-defined context of formal systems. For a system to be called a formal system it must first be capable of translating any statement it concerns itself with to a formula in the

[4] For a more complete treatment see my (Barcsák 2017). The third of these similarities seems to me to be the most important one and I will rely on this in section 5 of this paper.

notation of the system; that is, to a string of symbols which is then manipulated by the system in a totally mechanical[5] way without regard to the meanings we originally attributed to the symbols. For in such a system, as Gödel puts it, "the meaning of the symbols is immaterial, and it is desirable that they be forgotten" (Gödel 1965, 153). This is precisely why such systems are called *formal*: the manipulations the strings of symbols undergo are governed by mechanical rules which affect the strings only on the basis of their form, totally disregarding their meanings. Typically, formal systems are built up by selecting a countable number of formulas, the axioms, and specifying the rules of manipulation in formal (or syntactic) terms.[6]

What counts as a formal system has been very clearly determined as a result of the 20th-century development of formal logic. It has been clarified, in particular, what we can consider to be an entirely mechanical system (that is, formal in the above sense).[7] It turns out that there is a class of formal systems which are mutually translatable into each other and hence equivalent, which represent everything that we can be certain is fully mechanically, formally representable. There are simpler formal systems that express less than this class of systems, but such simpler systems are fully represented in the latter; and there are more complex systems which express more than this class of systems, but which are not fully mechanical/formal. This class of systems, therefore, comprises everything that we now know is mechanically controllable.[8] In what follows, I will rely on one formalization of this kind of system, Douglas Hofstaedter's Typographical Number Theory (TNT) (Hofstadter 1979).

Such systems can express a great deal. TNT, for example, was designed to capture everything that we know about natural numbers and their relations. It can thus formalize any statement about natural numbers: statements such as 7+2=9, or

[5] By "mechanical" I mean representable by a Turing-machine. In this sense, "mechanical" is synonymous with "effectively calculable" or with "reducible to a computable function of integers" (Gödel 1995, p. 304n1).

[6] David Hilbert, the initiator and main advocate of formalism in mathematics, describes formal systems as follows:

> We now divest the logical signs of all meaning, just as we did the mathematical ones, and declare that the formulas of the logical calculus do not mean anything in themselves… In this way we now finally obtain, in place of the contentual mathematical science that is communicated by means of ordinary language, an inventory of formulas that are formed from mathematical and logical signs and follow each other according to definite rules. Certain of these formulas correspond to the mathematical axioms, and to contentual inference there correspond the rules according to which the formulas follow each other; hence contentual inference is replaced by manipulation of signs according to rules, and in this way the full transition from a naïve to a formal treatment is now accomplished. (Hilbert 1967, 381)

[7] A useful summary of the events that led to this realization – and in particular of the effects of Alan Turing's paper (Turing 1936) – is provided by Juliette Kennedy (Kennedy 2014, 114–119).

[8] It is important to emphasize that this is just what we *know* is mechanically controllable. We know that what can be captured in a Turing machine is mechanically controllable. On the other hand, the reverse claim – that is, that everything that is mechanically controllable is captured in a Turing-machine – cannot be proved. This is usually referred to as the Church-Turing thesis, and we know that – in addition to Alonso Church and Alan Turing – Gödel also believed that this thesis holds.

3×5≠10, which are true, but also false statements such as 2×2=6. These statements will look as follows in TNT notation:

$$\mathrm{SSSSSSSO}+\mathrm{SSO}=\mathrm{SSSSSSSSSO}$$
$$\mathrm{SSSO}\cdot\mathrm{SSSSSO}\neq\mathrm{SSSSSSSSSSO}$$
$$\mathrm{SSO}\cdot\mathrm{SSO}=\mathrm{SSSSSSO}$$

These are very simple statements, but TNT can express much more complex assertions about numbers, too: for example, it can formalize statements such as "there are infinitely many prime numbers," or "the expression $x^n+y^n=z^n$ has no integer solutions for $n>2$," or "every even number that can be expressed as the sum of two primes." The first of these is translated into TNT notation in this way: $\forall d: \exists e: \sim\exists b: \exists c: (d + Se) = (SSb \cdot SSc)$,[9] and for the other two a similar TNT translation is also possible. In fact, TNT is complex and expressive enough to formalize potentially any statement about natural numbers.

What is more, it can even produce a complete list of all the meaningful arithmetical statements by means of formalization. It can rule out in a completely mechanical way all meaningless statements, such as for example × 12 + −6 = 66. A statement like this obviously does not make sense because it is not well formed (it just does not use the symbols in the right way), and TNT can always determine by a mechanical procedure whether or not any statement expressed in its notation is well-formed. As a result, we can select only the well-formed formulas (wff) of the system. Moreover, we could even organize these into a list, for example, on the basis of the length of the formulas, starting with the shortest and moving towards increasingly longer ones. Among formulas of equal length, we could create order by some alphabetization, and in this way, in theory, we could compile the complete list of well-formed formulas. This list would of course be an infinite one, but it is countably infinite, which means that we can even number the formulas, assigning a unique natural number (of which there are likewise an infinite number) to each item on the list.

Another important property of formal systems is that not only are they capable of producing a complete list of all the well-formed formulas, but they can also enumerate all the *theorems* of the system; that is, all the formulas that can be derived from the axioms by the mechanical application of the rules of procedure. In other

[9] Where b, c, e and d are integer variables, ∀ and ∃ are the usual universal and existential quantifiers ("for all" and "there exists", respectively), ~ is the negation operator, and S represents the successor function ("successor of").

words, formal systems are also capable of generating a complete list of the formulas that they can prove (that is, formally derive from the axioms).

What formal systems can provide is thus two lists: one containing all the possible well-formed formulas (that is, all the meaningful statements) about numbers, and the other comprising all the provable formulas. With the help of our formal system, therefore, we can potentially reduce the question of arithmetical truth to a mechanically controllable procedure. We take a random formula from the list of well-formed formulas – say "2×2=6" or "the expression $x^n+y^n=z^n$ has no integer solutions for n>2" – and ask, "Is this on the list of theorems?" If it is, then it is true, and if not, then it is false. To ascertain that an arbitrary well-formed formula is on the list of theorems we must demonstrate that the given formula can be gained by the mechanical manipulation (that is, a purely formal, syntactic handling) of the formulas representing the axioms. This is what is called a proof. Sometimes it is easy to prove whether a well-formed formula is on the list of theorems. In just a few steps, for example, we could prove that the formula representing "2×2=6" is not on the list; at other times the proof is rather more complicated. It took more than three and a half centuries to prove that Fermat's last theorem ("the expression $x^n+y^n=z^n$ has no integer solutions for n>2") is on the list, and the demonstration is more than 120 pages long (Andrew Wiles proved it in 1994–95) (Wiles 1995). We still do not know whether the statement "every even number can be expressed as the sum of two primes" is on the list – it probably is, because this is Goldbach's conjecture, which is in all likelihood true, but ever since the conjecture was first formulated in 1742, no one has succeeded in demonstrating it. In principle, however, we could expect that such a demonstration may eventually be carried out and that thus the truth of arithmetical propositions can always be determined entirely mechanically.

3. Reference and Truth

Once we establish this, however, the question arises: "In what sense could the theorems of a formal system like this be said to be 'true'?" What we mean by "true" is generally the so-called correspondence conception of truth; that is, the view under which – to use Alfred Tarski's phrase – "[t]he truth of a sentence consists in its agreement with (or correspondence to) reality" (Tarski 1944, 343) or, to use another formulation by the same author, "[a] sentence is true if it designates an existing state of affairs" (Tarski 1944, 343). But if the system producing the theorems is fully mechanical, then how can we know that the formulas mechanically produced actually correspond to states of affairs in an objective reality? If the system is purely mechanical, then there is a chance that everything it produces is mere tautology and that all that its operations amount to is merely, as Gödel puts it, "an idle running of language" (Gödel 1995, 319).[10]

This is the question of the reference of formal systems – it is a vast topic in the philosophy of mathematics and I will not be able to go into the details here. Suffice it to say that in my book I come to the conclusion that for a totally mechanically conceived, purely formally or syntactically specified system we must minimally presuppose two things to be able to maintain that the system is referential, that is, that it can sustain the correspondence conception of truth: we must presuppose (1) that the system is consistent, and (2) that it contains true but unprovable sentences – that is, well-formed formulas that represent truths, though we cannot derive them as theorems.[11]

The first of these conditions is relatively easy to justify. By the laws of classical logic,[12] out of a formal contradiction everything follows (ex contradictione quodlibet sequitur – ECQ). This means that if our system were inconsistent – that is, if it could prove a contradiction – then it would prove every formula. If every formula were true, then truth obviously could not be used in the correspondence sense; it just would not make sense to maintain that every state of affairs exists at the same time. It is therefore clear that the formal consistency of the system is an indispensable precondition for formulating any notion of truth in the correspondence sense.

Unlike this first condition, however, the second – that is, that the system should contain true but unprovable sentences – is thoroughly problematic. For starters, as Tarski proved, the concept of truth cannot be formulated inside a given formal system.[13] Consequently, and secondly, if we assume true but unprovable sentences, we beg the question, that is, we assume that we know what truth is before we could

[10] This situation is closely analogous to the philosophical problem usually referred to as the "paradox of analysis." This paradox was first pointed out by G. E. Moore and received its classic formulation from C. H. Langford, which runs thus:

> Let us call what is to be analyzed the analysandum, and let us call that which does the analysing the analysans. The analysis then states an appropriate relation of equivalence between the analysandum and the analysans. And the paradox of analysis is to the effect that, if the verbal expression representing the analysandum has the same meaning as the verbal expression representing the analysans, the analysis states a bare identity and is trivial; but if the two verbal expressions do not have the same meaning, the analysis is incorrect. (Langford 1968, 323) Cf. also (Norris 2012, 141).

[11] The first of these requirements is intuitively obvious. The second can be formulated in several different ways. That all these different ways can be summarized and succinctly stated in this one requirement of the presence of true but unprovable sentences is something that I arrived at as a result of an analysis of Tarski's invocation of the principle of the excluded middle in his (Tarski 1983).

[12] By "classical logic" I simply mean the standard logic of mathematical practice (by and large the propositional and predicate calculuses), as distinct from, for example, intuitionistic logic or paraconsistent logics.

[13] This is what is usually referred to as "Tarski's Theorem" and he first presented it in the "Postscript" to his (Tarski 1983, 268–277).

ground this concept. Thirdly, and for us most importantly, this would involve a naïve presupposition of the independent, objective existence of reality: for us to know that the true but unprovable sentence is true, we would need to assume that we have access to the state of affairs the sentence refers to before formulating this knowledge in the sentence itself.

The first two of these consequences seem to me to be inevitable: since the concept of truth cannot be expressed in a consistent formal system, any account of reference will to some extent beg the question. For any such account we will need to assume an external point of view, we will have to presuppose at least the possibility of reference. But does this mean that we likewise need a naïve presupposition of objective existence? Not necessarily. One of the central claims of my book is precisely this: that we *can* ground reference for formal systems without presupposing an objective reality. But for this we need first Gödel's insight about the inevitable presence of undecidable sentences in formal systems, and secondly, Derrida's insight about the role of this undecidability in grounding the possibility of reference.

4. The Gödelian Insight

Let us examine these insights one by one. What Gödel showed in his famous 1931 paper "On Formally Undecidable Propositions Of *Principia Mathematica* And Related Systems" (Gödel 1992) is that – although we cannot establish true but unprovable sentences inside a formal system – we *can* always produce undecidable formulas inside such a system on strictly formal grounds. He demonstrated this in two steps, both of which required remarkable genius and neither of which will I be able to represent in any depth, so I am just giving a sketch of Gödel's procedure:[14]

First, he proved that statements *about* the formal system can be translated into statements *in* the formal system. Thus, statements such as "formula x has a proof in the system" can be directly transformed into well-formed formulas of the system itself. He showed, in other words, that formal systems are capable of reflecting their own operations, that they can represent their own syntax.

Second, he showed how we can formulate an undecidable sentence on this basis. As illustration, consider the sentence "the nth well-formed formula is not on the list of TNT theorems." This is a clear and unambiguous statement about the functioning

[14] Several accessible accounts of Gödel's procedure are now available, such as (Hofstadter 1979) (Berto 2009) (Smullyan 1992) (Franzén 2005, 10–57) (Wright 1994, 185-186). In what follows I will adapt – and further simplify – Roger Penrose's simple but elegant account in *The Emperor's New Mind* (Penrose 1989, 138–141).

of TNT, so – on the basis of the first point above – we can formulate it as a well-formed formula of TNT itself. As such it will be listed among the well-formed formulas of the system (such a listing, as we have seen, is always possible) and will be assigned a unique number: the kth well-formed formula, say. Now n is a free variable in our formula, which means that it can be replaced by any concrete natural number. This formula will, therefore, give rise to an infinite family of formulas: "the first well-formed formula is not on the list", "the second well-formed formula is not on the list", etc. In the case of each of these formulas we can check if what they state is actually true or not. We can seek a proof for the first well-formed formula, then for the second, and so on. In each case, we will in theory be able to determine if the given formula has a proof inside the system or not. But what happens if we come to the kth formula on the list and substitute k for n? This will be a perfectly legitimate formula, just like any other on the list of well-formed formulas. However, it will make a statement, curiously enough, about itself. It will state, to be precise, that it is not on the list of theorems.[15] Will this formula then be on the list of theorems? If it is, then we will end up with a contradiction, for what the formula states is precisely that it is not on the list. If, on the other hand, it is not on the list, then – by a law of formal logic – its negation must be on the list, which asserts that the original formula is on the list, and this will again lead us to a contradiction. This means that neither the formula itself nor its negation can be on the list of theorems – assuming only that the formal system is consistent. This formula, in other words, will be neither provable, nor disprovable: it will be *undecidable*.

This is of course a rather drastically simplified and not even entirely consistent demonstration of Gödel's procedure, but the idea relevant for us here is that Gödel could demonstrate beyond doubt that in any formal system of the type we are discussing here there will always be such undecidable formulas. How does this modify the situation in regard to our ambition to ground the reference of formal systems? Remember that for establishing the correspondence conception of truth we need – apart from assuming the consistency of the system – true but unprovable formulas. Since by Tarski's Theorem we cannot capture the concept of truth inside the system, we do not seem to be much better off now that we have established the existence

[15] Penrose establishes this by first pointing out that a list of all propositional functions that depend on a single variable can in principle be compiled. Then he shows that the propositional function that asserts that the nth propositional function on this list has no proof in the system is a propositional function that depends on a single variable and must therefore be included in the list comprised of all such propositional functions. This means that it must have a unique ordinal number assigned to it, say it is the kth propositional function on the list. Finally, Penrose obtains the Gödel sentence by substituting k for n, which results in the kth propositional function asserting about itself that it has no proof in the system. (Penrose 1989, 138–140)

of undecidable formulas. Undecidable formulas are certainly unprovable, but why should they be true? This question cannot be answered in a fully convincing way. Assuming the truth of undecidable sentences will always remain just an assumption, which we need in order to account for the reference of formal systems.

However, there is a sense in which we are still somewhat better off once we have undecidable sentences. For if we have undecidable sentences, then it is clear that we *can* assume the existence of true but unprovable sentences. We can do so simply because the undecidable formula is clearly beyond what the system can mechanically control: since it is undecidable, it is clearly unprovable and as such it *could* be true for all we know. There is no way we can formally prove the contrary by means of our formal system. What is more, with this conception in mind we become capable of developing a new sense of the truth of true but unprovable sentences (which we must illegitimately assume anyway), a sense which does not require presupposing objective existence. It is for this step that we need the Derridean insight. Let us see how.

5. The Derridean Insight

So, what does the truth of the undecidable sentence (or of its negation)[16] mean if we choose to assume it to be true? The intuitive interpretation is of course that it means that it is true in the correspondence sense – that is, by virtue of referring to an objective state of affairs which exists. This was actually Gödel's own interpretation, too: if we have two contradictory sentences such that one is the negation of the other, we must conclude that one of them is true. In the case of the undecidable sentence, we know furthermore that neither it nor its negation can be proved, and this leads directly to the conclusion that there are truths that simply cannot be captured by the formal system. If we interpret truth here in the correspondence sense, then this means that there are certain states of affairs which our system just cannot grasp. No matter how we set up a formal system, the reality that it refers to will always exceed the capacities of the system: it will always be in excess of whatever system we design to refer to it.[17] For Gödel, therefore, the inevitable presence of

[16] If we view the undecidable formula simply as a syntactic construction, the assumption of its truth is just as valid as the assumption of the truth of its negation, since the requirement of consistency only demands that they must not both be true at the same time. In what follows I will only talk about assuming the truth of the undecidable formula itself, but the argument can also apply – with some complications that I will not go into here – if we assume the truth of its negation.

[17] Gödel was of course more subtle than this when formulating his position (cf. especially his (Gödel 1995a) and (Gödel 1995b)). Nonetheless, he was a mathematical Platonist, meaning that he believed in the independent existence of an objective mathematical reality beyond that which can be grasped in formal systems. For an account of Gödel which emphasizes this realist streak in his thought see (Goldstein 2005).

undecidable formulas marks a fundamental incapacity of any formal system, the impossibility of grasping reality, or even a well-defined segment of it, in its entirety. This is also expressed in the name of the theorems he based on his demonstration of the existence of formally undecidable sentences: these are called the *incompleteness* theorems, implying that any formal system is incomplete in the sense that it cannot prove all truths about the reality it describes.[18]

Must we, however, interpret the inevitable presence of undecidable sentences as a limitation? One of the central insights of deconstruction is that we do not need to. For we can also consider such limitations, such impossibilities, as necessary conditions for a possibility. As Giorgio Agamben puts it in "Pardes," his homage to Derrida:

> It does not suffice, however, to underline (on the basis of Gödel's theorem) the necessary relation between a determinate axiomatics and undecidable propositions: what is decisive is solely how one conceives this relation. It is possible to consider an undecidable as a purely negative limit (Kant's *Schranke*), such that one then invokes strategies (Bertrand Russell's theory of types or Alfred Tarski's metalanguage) to avoid running up against it. Or one can consider it as a *threshold* (Kant's *Grenze*), which opens onto an exteriority and transforms and dislocates all the elements of the system. (Agamben 1999, 214)

Agamben's point is of course that deconstruction follows the second path. Derrida's undecidables, such as the hymen, the trace, the supplement, the gift, hospitality, etc., are thresholds. Naturally, they mark a fundamental impossibility, but an impossibility which is also the condition of the possibility of that which they render impossible. And I think we can use this insight to reinterpret the function of the inevitable undecidable formula in any formal system. In particular, we can interpret the impossibility of reference marked by Gödel's undecidable formula in a given formal system as the condition of the possibility for this system to be referential at all, to make reference to something other than itself. We have seen that without undecidable sentences, formal systems cannot be considered referential: they cannot refer to anything other than themselves, since they are only capable of exhibiting "an idle running of language." With undecidable sentences, however, it becomes possible to assume the truth of these sentences and thus we become capable of accounting for reference. The undecidable sentence itself is of course a

[18] This is, incidentally, the line the famous "Gödelian arguments" of John Lucas, Roger Penrose, and Stanley Jáki also take (Lucas 1961) (Lucas 1996) (Penrose 1989) (Penrose 1994) (Jáki 1966) (Jáki 2004).

point at which the functioning of the system breaks down, thus marking a point where reference is certainly impossible. As such, however, it provides a ground for assuming the possibility of reference. In fact, it is alone capable of establishing that a formal system can be more than just "an idle running of language"; it can alone guarantee that we can think of the other sentences of the system as referential in the correspondence sense; that is, as being made true or false according to the existence or non-existence of certain objective facts.

For this, however, it is not enough to have undecidable sentences. We must also assume the undecidable sentence to be *true*, and this brings us back to the original question: in what sense can the undecidable sentence be assumed to be true? We have seen that assuming its truth in the correspondence sense leads directly to Gödel's Platonism, to the excess of reality over the system, and thus inevitably to a naïve presupposition of the objective existence of reality. This, however, is not the only possible interpretation of the truth of undecidable sentences. For – and this is another Derridean insight – the undecidable can also be interpreted as marking – as Derrida puts it in relation to the hymen in *The Double Session* – "the irreducible excess of the syntactic over the semantic" (Derrida 1981b, 230).

That this possibility is indeed available becomes clear if we examine the situation arising from the requirement of true but unprovable sentences. For, as we have seen, we need true but unprovable sentences to be able to ground reference for a formal system in the correspondence sense. This means that – since we cannot establish the existence of such sentences by a formal proof – we must *presuppose* them *before* establishing the correspondence conception of truth. Therefore, the truth of the undecidable sentences is a *precondition* of this conception and does not need to be bound by it. Assuming that the undecidable sentence is true in the correspondence sense can at best be a retrospective projection of a sense of "true" that can only be established *after* we have presupposed the truth of undecidable sentences. Therefore, while it is true that the correspondence conception of truth depends on and is determined by the truth of the undecidable formula, the sense in which the latter is true need not be determined by the former.

The question that remains to be asked is "Can the undecidable formula (or its negation) be assumed to be true in any sense other than correspondence?" And this is where the Derridean insight cited above can again come to our assistance. For it highlights the possibility that the undecidable formula can be seen as referring solely to the syntactic system itself. If we interpret the truth of the undecidable formula in this way, then it will be true not of some preexisting, independent and objective reality, but of the formal system itself as a referential system. For if we do not presuppose the objective existence of reality, then the truth of the undecidable formula will simply mean that in any system we set up to refer to some reality there will always be formulas that must be true *regardless of* how things are in reality. The truth of such a formula will therefore depend not on an objective, independently existing state of affairs, but only on the formal requirements of our system, only on its syntax. The truth of undecidable formulas will thus attest to the independence of the system from any reality and will mark the excess of the syntactic system over whatever reality it refers to.[19]

Relying on this sense of the truth of the undecidable formula we become capable of grounding an account of truth as correspondence between a formal system and reality, or, in other words, we become capable of accounting for the reference of formal systems. What is more, we become capable of doing this without a naïve presupposition of objective existence. For by exhibiting a formula whose truth certainly does not depend on any objective existence, we can establish the independence of the formal system, its autonomy from any reality that it may refer to. We can establish, in other words, that the formal system is not reality. And once we have thus established the independence of the system, we become capable of assuming that it is independent *from something other than itself*. In this way, therefore, what seemed to be an incapacity, an impossibility in the formal system, turns out to be the ultimate condition of the possibility of grounding its reference. Because the system can be thought of as independent, we *can* think that it is related to something entirely other than itself.

References

Agamben, Giorgio. 1999. "Pardes: The writing of potentiality." In *Potentialities: Collected Essays in Philosophy*, by Giorgio Agamben, edited by Daniel Heller-Roazen, translated by Daniel Heller-Roazen, 205–219. Stanford, CA: Stanford University Press. https://doi.org/10.1515/9780804764070

Barcsák, János V. 2017. "Formalization, Politics, Creativity." In *Intertextuality, Intersubjectivity, and Narrative Identity*, edited by Péter Gaál-Szabó, 5-20. Newcastle upon Tyne: Cambridge Scholars.

[19] Whether such an interpretation makes any sense in the context of formal systems and their reference is of course not self-evident, and one of the major tasks I had to face in my book was to develop the precise sense in which this view of the truth of the undecidable formula can be meaningful in such contexts as, for example, the truth of arithmetical propositions.

Berto, Francesco. 2009. *There's Something about Gödel: The Complete Guide to the Incompleteness Theorems*. Wiley-Blackwell. https://doi.org/10.1002/9781444315028

Derrida, Jacques. 1981a. "Différance." In *Margins of Philosophy*, by Jacques Derrida, translated by Alan Bass, 1–27. Chicago: University of Chicago Press.

—. 1981b. "The Double Session." In *Dissemination*, by Jacque Derrida, translated by Barbara Johnson, 187-315. London; New York: Continuum.

—. 1988. "Afterword: Toward an Ethic of Discussion." In *Limited Inc*, by Jacques Derrida, 111-160. Evanston, IL: Northwestern University Press.

Franzén, Torkel. 2005. *Gödel's Theorem: An Incomplete Guide to Its Use and Abuse*. Wellesley, MA: A. K. Peters. https://doi.org/10.1201/b10700

Gödel, Kurt. 1965. "On Undecidable Propositions of Formal Mathematical Systems." In *The Undecidable. Basic Papers On Undecidable Propositions, Unsolvable Problems And Computable Functions*, edited by Martin Davis, 39–74. Hewlett, NY: Raven Press.

—. 1992. *On Formally Undecidable Propositions Of Principia Mathematica And Related Systems*. New York: Dover Publications.

—. 1995a. "Is mathematics syntax of language? - III." In *Collected Works. Volume III. Unpublished Essays and Lectures*, by Kurt Gödel, edited by Solomon Feferman, John W. Dawson, Warren Goldfarb, Charles Parsons and Robert M. Solovay, 334–356. New York, Oxford: Oxford University Press.

—. 1995b. "Some basic theorems on the foundations of mathematics and their implications." In *Collected Works. Volume III. Unpublished Essays and Lectures*, by Kurt Gödel, edited by Solomon Feferman, John W. Dawson, Warren Goldfarb, Charles Parsons and Robert M. Solovay, 304–323. New York, Oxford: Oxford University Press.

Goldstein, Rebecca. 2005. *Incompleteness: The Proof and Paradox of Kurt Gödel*. New York, London: Atlas, Norton.

Hilbert, David. 1967. "On the Infinite." In *From Frege to Gödel: A Source Book in Mathematical Logic 1879–1931*, edited by Jean van Heijenoort, 367–392. Cambridge, MA: Harvard University Press.

Hofstadter, Douglas R. 1979. *Gödel, Escher, Bach: an Eternal Golden Braid*. London: Penguin Books.

Jáki, Stanley L. 1966. *The Relevance of Physics*. Chicago: University of Chicago Press.

—. 2004. "A Late Awakening to Gödel in Physics." *Sensus communis* 5 (2–3): 153–162.

Kennedy, Juliette. 2014. "Gödel's 1946 Princeton bicentennial lecture: an appreciation." In *Interpreting Gödel: Critical Essays*, edited by Juliette Kennedy, 109–130. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511756306.009

Langford, Cooper Harold. 1968. "The Notion of Analysis in Moore's Philosophy." In *The Philosophy of G. E. Moore*, edited by Paul Arthur Schilpp, 321–341. La Salle, IL: Open Court.

Livingston, Paul M. 2010. "Derrida and Formal Logic: formalizing the undecidable." *Derrida Today* 3 (2): 221–239. https://doi.org/10.3366/drt.2010.0205

—. 2012. *The Politics of Logic. Badiou, Wittgenstein, and the Consequences of Formalism*. New York; London: Routledge.

Lucas, John R. 1961. "Minds, Machines, and Gödel." *Philosophy* 36: 112-127. https://doi.org/10.1017/S0031819100057983

—. 1996. *Minds, Machines, and Gödel: A Retrospect.* Vol. 1, in *Machines and Thought. The Legacy of Alan Turing*, edited by Peter Millican and Andy Clark, 103–124. Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198235934.003.0007

Norris, Christopher. 2006. "Deconstruction, Analysis, and Deviant Logic: Derrida at the Limits of Thought." *The Harvard Review of Philosophy* 36-61. https://doi.org/10.5840/harvardreview20061413

—. 2012. *Derrida, Badiou and the Formal Imperative.* London; New York: Continuum. https://doi.org/10.5040/9781350251816

Penrose, Roger. 1994. *Shadows of the Mind.* Oxford: Oxford University Press.

—. 1989. *The Emperor's New Mind.* Oxford: Oxford University Press.

Priest, Graham. 1994. "Derrida and self-reference." *Australasian Journal of Philosophy* 72 (1): 103-114. https://doi.org/10.1080/00048409412345911

—. 2002. *Beyond the Limits of Thought.* Oxford: Clarendon Press. https://doi.org/10.1093/acprof:oso/9780199254057.001.0001

—. 2006. *In Contradiction: A Study of the Transconsistent.* Oxford: Clarendon Press. https://doi.org/10.1093/acprof:oso/9780199263301.003.0015

Smullyan, Raymond M. 1992. *Gödel's Incompleteness Theorems.* Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780195046724.001.0001

Tarski, Alfred. 1944. "The Semantic Conception of Truth: and the Foundations of Semantics." *Philosophy and Phenomenological Research* (International Phenomenological Society) 4 (3): 341–376. https://doi.org/10.2307/2102968

—. 1983. "The Concept of Truth in Formalized Languages." In *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*, by Alfred Tarski, edited by John Corcoran, translated by J. H. Woodger, 152–278. Indianapolis: Hackett Publishing.

Turing, Alan M. 1936. “On Computable Numbers, with an Application to the Entscheidungsproblem.” *Proceedings of the London Mathematical Society* s2–42 (1): 230–265. https://doi.org/10.1112/plms/s2-42.1.230

Wiles, Andrew John. 1995. “Modular Elliptic Curves and Fermat’s Last Theorem.” *Annals of Mathematics* 141: 443–551. https://doi.org/10.2307/2118559

Wright, Crispin. 1994. “About "The Philosophical Significance of Gödel's Theorem": Some Issues.” In *The Philosophy of Michael Dummett*, edited by Brian McGuiness and Gianluigi Oliveri, 167–202. Dordrecht; Boston; London: Kluwer Academic Publishers. https://doi.org/10.1007/978-94-015-8336-7_9