The Review of Symbolic Logic
Volume 16, Number 2, June 2023

GÖDEL'S THEOREM AND DIRECT SELF-REFERENCE

SAUL A. KRIPKE

Abstract. In his paper on the incompleteness theorems, Gödel seemed to say that a direct way of constructing a formula that says of itself that it is unprovable might involve a faulty circularity. In this note, it is proved that 'direct' self-reference can actually be used to prove his result.

In his epochal 1931 paper on the incompleteness theorems Gödel wrote, regarding his undecidable statement, saying, in effect, "I am unprovable,"

> Contrary to appearances, such a proposition involves no faulty circularity, for initially it [only] asserts that a certain well-defined formula (namely, the one obtained from the qth formula in the lexicographic order by a certain substitution) is unprovable. Only subsequently (and so to speak by chance) does it turn out that this formula is precisely the one by which the proposition itself was expressed. [1, p. 151, fn. 15]

The tendency of this footnote might be that if one tried to express the statement directly, there might indeed be a faulty circularity. In my own paper, "Outline of a Theory of Truth," I expressed some doubt about this tendency, and stated that the Gödel theorem could be proved using 'direct' self-reference. To quote this paper,

> A simpler, and more direct, form of self-reference uses demonstratives or proper names: Let 'Jack' be a name of the sentence 'Jack is short', and we have a sentence that says of itself that it is short. I can see nothing wrong with "direct" self-reference of this type. If 'Jack' is not already a name in the language,[4] why can we not introduce it as a name of any entity we please? In particular, why can it not be a name of the (uninterpreted) finite sequence of marks 'Jack is short'? (Would it be permissible to call this sequence of marks "Harry," but not "Jack"? Surely prohibitions on naming are arbitrary here.) There is no vicious circle in our procedure, since we need not *interpret* the sequence of marks 'Jack is short' before we name it. Yet if we name it "Jack," it at once becomes meaningful and true. (Note that I am speaking of self-referential sentences, not self-referential propositions.[5])

Received: October 27, 2020.
2020 *Mathematics Subject Classification*: 03-03, 3B10.
Key words and phrases: direct self-reference, Gödel numbering, Incompleteness.

doi:10.1017/S1755020321000526

In a longer version, I would buttress the conclusion of the preceding paragraph not only by a more detailed philosophical exposition, but also by a mathematical demonstration that the simple kind of self-reference exemplified by the "Jack is short" example could actually be used to prove the Gödel incompleteness theorem itself (and also, the Gödel–Tarski theorem on the undefinability of truth). Such a presentation of the proof of the Gödel theorem might be more perspicuous to the beginner than is the usual one. It also dispels the impression that Gödel was forced to replace direct self-reference by a more circumlocutory device. The argument must be omitted from this outline.[6] [6, pp. 77–78].

[4] We assume that 'is short' *is* already in the language.

[5] It is *not* obviously possible to apply this technique to obtain "directly" self-referential *propositions*.

[6] There are several ways of doing it, using either a nonstandard Gödel numbering where statements can contain numerals designating their own Gödel numbers, or a standard Gödel numbering, plus added constants of the type of 'Jack'.

As I say in footnote 6, there are several ways of obtaining the Gödel theorem using direct self-reference, analogously to 'Jack is short'. Note that, as I say in footnote 5, I am not claiming that this technique could be used to obtain self-referential *propositions*. Perhaps this is what Gödel had in mind in his own footnote. Let us look at how Gödel's first incompleteness theorem is to be done these ways.

One way, observed independently by Raymond Smullyan, is to use a nonstandard Gödel numbering where a formula can contain a numeral designating its own Gödel number.[1] This is clearly impossible under the usual Gödel prime power numbering, or various other variants. That this should not happen might be a natural restriction, since a Gödel number might be thought to correspond with a formula as a composite object, and this might be thought to preclude a formula from containing a numeral designating its own Gödel number.

Nevertheless we propose a nonstandard Gödel numbering allowing a statement to contain a numeral designating its own Gödel number. We can proceed as follows. Let x_1 be a fixed variable. Let $A_1(x_1)$, $A_2(x_1)$, ...be an enumeration of all those formulae that contain at most x_1 free (Gödel's class signs). We assume that the language of the system studied contains, either directly or by virtue of interpretation, the language of arithmetic. Numerals can be assumed to be terms 0 followed by (or preceded by) finitely many successor symbols (allowing none). We use $0^{(n)}$ for 0 with n successor symbols; $n = 0$ is allowed, as I said. $0^{(n)}$ denotes the number n.

Let the 'original' Gödel numbering be Gödel's own prime power product numbering, except that the smallest prime used is 3, so that Gödel numbers are always odd. In the 'new' numbering, all Gödel numbers coincide with the 'original', except that for each n, the formula

$$(\exists x_1)(x_1 = 0^{(2k_n)} \wedge A_n(x_1)),$$

[1] Almost everything I say about Smullyan is based on conversation with him. We haven't been able to find a published reference.

https://doi.org/10.1017/S1755020321000526 Published online by Cambridge University Press

gets the Gödel number $2k_n$, where k_n is the original Gödel number of $(\exists x_1)(x_1 = 0^{(n)} \wedge A_n(x_1))$. The 'new' numbering allows a formula to contain a numeral designating its Gödel number, and in that sense it is a self-referential Gödel numbering.

In this self-referential Gödel numbering, every formula $A_n(x_1)$ has an 'instance' $(\exists x_1)(x_1 = 0^{(2k_n)} \wedge A_n(x_1))$ asserting that its own Gödel number satisfies $A_n(x_1)$. The Gödel incompleteness theorem is the special case where $A_n(x_1)$ is unprovability in the system.

The 'new' Gödel number is in effective 1 – 1 correspondence with the old one, where the inverse function is also effective. The complexity of a property in the arithmetical hierarchy (Σ^0_1, Π^0_1, etc.) does not change when the 'new' Gödel numbers replace the old.[2]

Remark 1. We could have simply used $A_n(0^{(2k_n)})$. But using $(\exists x_1)(x_1 = 0^{(2k_n)} \wedge A_n(x_1))$ has two advantages, both emphasized by Raymond Smullyan. First, concatenation is easier to arithmetize than substitution, which has some value even here.[3] Second, this choice guarantees that no formula gets two Gödel numbers, which is not clear if we use the simpler version. It also may not be harmful if some formulae get two or more Gödel numbers, but in that case we cannot think of formulae as *identified* with their Gödel numbers.

Remark 2. A referee has proposed an alternative and simpler 'new' Gödel numbering. For each n, and formula $A(x_1)$, let $(\exists x_1)(x_1 = 0^{(2n)} \wedge A_n(x_1))$ get the Gödel number $2n$. Otherwise, the 'old' (always odd) numbers are used.

Much more natural, in my opinion, for getting 'direct' self-reference, is the use of added constants.[4] Once again, let $A_1(x_1), A_2(x_1), \dots, A_i(x_1)$ be an enumeration of all those formulae of the language L containing at most the variable x_1 free. Let $a_1, a_2, \dots$ be a denumerable list of constants, none of which are in the original language L. Give Gödel numbers to the formulae of the extended language in a conventional way. Now add to the axioms of the original system S, an infinite set of axioms $a_i = 0^{(n_i)}$, where n_i is the Gödel number of $A_i(a_i)$. The resulting system is S'. Pretty clearly S' is a conservative extension of S. It simply extends S by adding an infinite set of constants

[2] The existential quantifier in $(\exists x_1)(x_1 = 0^{(2k_n)} \wedge A_n(x_1))$ does not increase complexity, since the formula is equivalent to $A(0^{(2k_n)})$. For the Gödel theorem, the important feature is that 'unprovability in the system' is still a Π^0_1 predicate. If the language contains non-arithmetical predicates that have hierarchical status, say, in the hyperarithmetical or analytical hierarchy, these also remain the same.

Further, since the domain and range of both of these 1 – 1 mappings are recursive, the mapping can be extended to a recursive permutation of the natural numbers.

[3] The second reason is the most important one. I have not worried about this in the case of added constants. But substitution of a constant for a fixed free variable x_1 is in fact easy to arithmetize provided x_1 does not occur bound in $A(x_1)$—in other words, that '(x_1)' does not occur in it.

[4] One of the referees has commented, regarding numerals, "it is highly plausible to view them as proper names; but they have also been seen as analogues of quotational names and structural-descriptive names. Numerals—unlike constants—are complex closed terms and at least in this respect similar to a term built from a term with function symbols for the primitive recursive functions for substitution and numerals." The doubts in question do not apply to constants, as the referee says. Nevertheless, the use of a self-referential Gödel numbering, though somewhat artificial, does seem more directly self-referential than the traditional Gödelian argument.

with specific numerical values—new names for particular numbers. Every proof in S' becomes a proof in S if each constant is replaced by the corresponding numeral.

Now, for the Gödel (first incompleteness) theorem, consider the predicate (in the original language, without the extra constants) that says that x_1 is the Gödel number of a formula unprovable in the extended system S'. Call this $\sim\mathrm{Thm}_{S'}(x_1)$. This is a Π^0_1 predicate and is a formula $A_i(x_1)$ for some i. In our construction, there is a constant a_i of S' such that $a_i = 0^{(n_i)}$ is an axiom of S', where n_i is the Gödel number of $A_i(a_i)$. Then it is easily shown, in the usual way, that $A_i(a_i)$ is unprovable in S' if S' is consistent, and that $\sim A_i(a_i)$ is unprovable in S' if S' is ω-consistent (or even 1-consistent). Given that S' is a conservative extension of S, the same things must be true of $A_i(0^{(n_i)})$ in the original system S as well.

The particular case of unprovability is just one example. For any definable property we obviously can use this construction to formulate a statement that 'says of itself' that it has that property (the 'self-reference lemma').

A more delicate construction uses only a single constant a, and a single extra axiom, in defining the system S^* extending S. It will have only one extra axiom, which will be called $(*)$. In the case of Gödel's first incompleteness theorem, we wish to construct a statement saying that it itself is unprovable. So given the constant a, we consider the single formula standing for the statement

$$a \text{ is unprovable in the system } S^*, \text{ which is } S \text{ extended by adding } \ulcorner a = 0^{(a)} \urcorner \text{ as a single axiom} \qquad (*)$$

(Of course this is to be written out symbolically.) The constant a occurs in $(*)$ only as explicitly shown. Note that in $(*)$, the constant a is both used and mentioned. In a standard Gödel numbering of the extended language, $(*)$ has a particular Gödel number. Say it is n. Then let the extended system S^* be axiomatized by adding $a = 0^{(n)}$ as an axiom.

Given this extra axiom, $a = 0^{(n)}$, note that $(*)$ will say that it itself (i.e., its Gödel number) is unprovable in the system S^* itself. Now we can once again deduce that $(*)$ is unprovable in S^* if S^* is consistent, its negation is unprovable in S^* if S^* is ω-consistent, etc. Also once again S^* is a conservative extension of S, so that any proof in S is valid in S^* iff it was valid in S, just as in the previous argument with an infinity of constants. And once again, this allows us to transfer the result to the original system S. Clearly also the argument extends to results about any definable $A(x_1)$ in place of unprovability (the self-reference lemma).

Remark. In spite of my own use of the Quinean terminology of 'use' and 'mention' in the preceding paragraph, I am not following his practices of quotation and quasi-quotation. Many formal expressions in the paper are used autonymously. The corners in $(*)$ are used to denote Gödel numbers of formulae $\ulcorner a = 0^{(a)} \urcorner$, where a is a constant on the left, but on the right it is an as yet unspecified numeral (the constant a is used, rather than mentioned). In other words, '$0^{(a)}$' is not a term in the language, since it means 0 followed by a successors, and therefore what formula is mentioned in the corners is yet to be specified. In the whole formula in corners, a is 'mentioned'.[5]

[5] A referee has suggested that the reader would profit from the following further readings on direct self-reference: [2–5, 7, 8].

https://doi.org/10.1017/S1755020321000526 Published online by Cambridge University Press

Acknowledgments. My thanks to Yale Weiss for editorial support. Special thanks to Romina Padró, Panu Raatikainen, and Albert Visser for their comments and for encouraging me to write this paper. This paper has been completed with support from the Saul Kripke Center, at The City University of New York, Graduate Center.

BIBLIOGRAPHY

[1] Gödel, K. (1986). On formally undecidable propositions of Principia Mathematica and related systems I. In Feferman, S., et al., editors. *Kurt Gödel: Collected Works, Volume I*. New York: Oxford University Press, translated by Jean van Heijenoort, pp. 145–195.

[2] Grabmayr, B., & Visser, A. (2021). Self-reference upfront: A study of self-referential Gödel numberings. *Review of Symbolic Logic*, doi: 10.1017/S1755020321000393.

[3] Halbach, V., & Visser, A. (2014). Self-reference in arithmetic I. *Review of Symbolic Logic* **7**, 671–691.

[4] ———. (2014). Self-reference in arithmetic II. *Review of Symbolic Logic* **7**, 692–712.

[5] Heck, R. (2004). Self-reference and the languages of arithmetic. *Philosophia Mathematica* **15**, 1–29.

[6] Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy* **72**, 696–716. Reprinted in Kripke, S. (2011). *Philosophical Troubles*. New York: Oxford University Press, pp. 75–98.

[7] Picollo, L. (2018). Reference in arithmetic. *Review of Symbolic Logic* **11**, 573–603.

[8] Visser, A. (2004). Semantics and the liar paradox. In Gabbay, D. and Guenthner, F., editors. *Handbook of Philosophical Logic*, Vol. 11 (second edition). Heidelberg: Springer, pp. 149–240.

THE SAUL KRIPKE CENTER, THE GRADUATE CENTER
CUNY, 365 FIFTH AVE., ROOM 7118
NEW YORK, NY 10016, USA
E-mail: kripkecenter@gc.cuny.edu