The Road to Gödel*

Saul A. Kripke

What does this title mean? Gödel's incompleteness theorem, as he originally presented it, seems to be an extraordinary and tricky magical construction. We can now, with modern recursion theory or computability theory,[1] reduce the problem to the unsolvability of the halting problem, or the theorem that there is a recursively enumerable set that is not recursive (computably enumerable set that is not computable) – though that is not the way Gödel presented it himself. A friend of mine of the highest distinction in this very field of recursion (computability) theory, once said to me in informal conversation that all of us know how the original Gödel statement was constructed, but no one really understands what it says. It is just an artificial product and has no intuitive content.

Now, I want to do two things: first, to present the Gödel theorem as almost the inevitable result of a historic line of thought. I don't mean that it *did* happen that way; I mean that it *could* have, and perhaps *should* have, but did not happen in that way. Second, I want to show that the Gödel statement, the one Gödel proves to be undecidable in the first incompleteness theorem, makes a fairly intelligible assertion that can actually be stated. Along the way I will do some things that are a little bit more technical, which are, not only as motivation but even as side theorems, unknown (as far as I know) even to specialists. They have to be separated out, because they involve more technicalities, though I will mention them and, to some extent, use them. I should also state at the outset that some of these issues are unnecessary to the main point about the interpretation of Gödel's statement, and can be avoided if one wishes to do so.

Let's look first at Gödel's own presentation, which has struck many readers as tricky and magical, by way of contrast with the present version. Gödel first says:

> Before going into details, we shall first sketch the main idea of the proof, of course without any claim to complete precision. The formulas of a formal system (we restrict ourselves here to the system *PM*) in outward appearance are finite sequences of primitive signs (variables, logical constants and parentheses or punctuation dots), and it is easy to state with

> complete precision *which* sequences of primitive signs are meaningful formulas and which are not.[2]
>
> … it does not matter what objects are chosen as primitive signs, and we shall assign natural numbers to this use. Consequently, a formula will be a finite sequence of natural numbers, and a proof array a finite sequence of finite sequences of natural numbers. (1931:147)

Now, Gödel gives his main construction. A formula with just one variable free, of the type of natural numbers, he calls a "class sign"; we can even restrict the notion so that the free variable is a particular one, x_1. The class signs are thought of as arranged in a series (sequence) and the n-th is denoted as $R(n)$. Both the concept "class sign" and the ordering relation R are definable in *Principia Mathematica* (*PM*). Any class sign $[\alpha; n]$ designates the formula obtained by replacing the free variable in the class sign α by the sign for the natural number n. The three-term relation $x = [y; z]$ is also definable in *PM*.

Then he defines a class of natural numbers:

$$n \in K \equiv \overline{Bew}\,[R(n); n]$$

This means that when we put in a numeral for the number n in place of the free variable in the nth class sign, which he calls $R(n)$, what we get is unprovable.[3] That itself defines a class, and so is denoted by some class sign. Hence, it is expressed by a class sign $R(q)$ for a certain natural number q. Then he shows that if we consider the proposition:

$$\overline{Bew}\,[R(q); q]$$

the result is undecidable, and in a tricky roundabout way can be interpreted as saying of itself that it is unprovable. He writes: 'the analogy of this argument with the Richard antinomy leaps to the eye. It is closely related to the 'Liar' too'. (1931:149). Pretty clever (if I don't give the details, I suppose you have heard them before) and pretty tricky construction. The result of his construction is that in an almost magical way it gives a formula uniquely applying to itself,

which can be proved to amount to its own provability. But what is the formula all about? What does it say? Gödel says in note 15 that the formula states only in an indirect way of itself that it is unprovable. He suggests that this shows the formula doesn't involve any faulty circularity, because only by accident does it turn out that it says of itself that it is unprovable.[4] It says that a formula obtained in a certain way is unprovable, and calculation shows that to be the original formula itself.

Unorthodoxly, instead of starting with any of the semantic paradoxes,[5] I want to start with what were ordinarily called "the paradoxes of set theory" in the old days – that is, that the so-called unrestricted comprehension schema leads to a contradiction.

The naïve or unrestricted comprehension schema is:

$$(\exists y)(x)(x \in y \equiv A(x))$$

For an arbitrary formula $A(x)$, as long as it does not contain y free, it defines a class of all those things where $A(x)$ holds. For those whose intuitions favored the schema, $A(x)$ could contain other parameters ($z_1, \ldots z_n$), and so the schema would read: ($z_1, \ldots z_n$) $(\exists y)(x)(x \in y \equiv A(x, z_1, \ldots z_n))$. However, since parameter-free versions already lead to paradox, we shall deal only with the version where $A(x)$ contains no parameters. Some of the classic paradoxes have versions that would be most obviously formalized using extensionality as well as comprehension. Even these can in fact be used to produce versions of the Gödel theorem, but there is no need for excessive complications.[6]

The contradictory nature of the schema shocked a great deal of the mathematical and logical community. Poincaré, whose attitude to both Cantor's set theory and mathematical logic was hostile, said that the Cantorians overreached themselves and ran into contradictions. Having earlier held that mathematical logic is sterile, now he exulted that it was no longer sterile; it leads to a contradiction (Poincaré 1912: 536-7). Cantor, however, never felt that the contradictory nature of the schema had anything to do with his set theory. I tend to believe that Hausdorff felt so also.[7] But Frege (1902) was certainly shocked. Russell (1967 [1902]), who popularized the paradoxes more than anyone else, was also shocked. Even those in the mathematical community who were more sympathetic to set theory than Poincaré thought there was a serious problem to be solved and that, naively, set theory was most naturally based on the inconsistent schema.[8]

Russell found the simplest and most famous counter-instance to the schema, where $A(x)$ can be a very simple formula, not even containing quantifiers. He showed that the case $x \notin x$ leads to a contradiction, because there can't be any y such that:

$$(x)(x \in y \equiv x \notin x)$$

Of course, as a special case of this, if we had such a y, we would derive:

$$y \in y \equiv y \notin y$$

This is only one of many consequences that are contradictory. Although Russell is really responsible for the recognition of the paradoxical character of other set theoretical paradoxes, such as the Burali-Forti paradox (see Moore and Garciadiego 1981), Quine (see 1951: 128-30) has found perhaps some of the simplest ones. For example, if $A(x)$ is:

$$\sim(\exists z)(\, x \in z \wedge z \in x \,)$$

a contradiction can be obtained too:

$$(\exists y)(x)[x \in y \equiv \sim(\exists z)(x \in z \wedge z \in x)]$$

This is the class of all unreciprocated classes. Let a class x be called *reciprocated* if it is a member of a class z which is also a member of x; *unreciprocated,* otherwise. Let K[9] be the class of all unreciprocated classes. Then is K itself reciprocated? Plainly not. For if some z reciprocates K

$$z \in K \wedge K \in z$$

since $z \in K$, z would have to be unreciprocated (by the definition of K). But it also would be reciprocated by K itself, which is a contradiction. So K is unreciprocated. Again by the definition of K, this means $K \in K$. But then

$$K \in K \wedge K \in K$$

i.e. K reciprocates itself, which is a contradiction. So this paradox also shows that the unrestricted comprehension axiom schema cannot be satisfied. Like Russell's paradox, it is far simpler than such classical set-theoretic paradoxes as the Burali-Forti paradox.

To me it has turned out to be of some conceptual interest that someone who was simply told that the unrestricted comprehension axiom schema is inconsistent, without needing any particular example, could already derive a (non-constructive) version of Gödel's theorem, proving under appropriate hypotheses (including consistency) that there is a true but unprovable universal statement, and which, on the assumption of ω-inconsistency, is undecidable. We shall prove this below.

Notice that what one has been told is independent of the interpretation of the epsilon relation. It is a matter of pure first-order logic that the unrestricted comprehension axiom schema is inconsistent. Russell already realized this in his example of the barber who shaves all and only those who do not shave themselves – and then the question is whether the barber shaves himself. Or, analogously to Quine's case, one could have the class of all barbers who are never shaved by any barber who shaves him also – or something like that. The same contradiction would go through, because the interpretation of the epsilon relation is irrelevant.[10]

Let us look at this matter in considerable generality. Consider a first-order structure having a domain D and some relations R_1, R_2, … that can have a certain number of places. Let's say that among these relations there is at least one two-place relation R_1. It could have, or it could not have, designated function letters and constants (whether these are allowed or eliminated[11] is not important):

$$<D, R_1, R_2, \ldots, f, \ldots>.$$

We have also a corresponding first-order language L with symbols – the variables range over the domain D, the relations are all primitive symbols of the language, and so on. Assume for the moment that D contains the formulas of its own language L in the domain (I will weaken that

assumption in a moment), or at least it contains the so-called Gödelian class signs involving one particular free variable:

$$\ulcorner A(x_1) \urcorner$$

(As I said one can normalize this by choosing a particular free variable.) Suppose we have a two-place predicate S between elements of the domain D and formulae:

$$aS\ulcorner A(x_1) \urcorner$$

(Here I am using "a" as an informal variable over elements of D. $\ulcorner A(x_1) \urcorner$ also ranges over elements of D, although only over particular elements, namely the 'class signs'.) Try to interpret this as:

$$a \text{ sat} \ulcorner A(x_1) \urcorner$$

It is an immediate consequence of the inconsistency of the unrestricted comprehension axiom that satisfaction by a class sign of the language – satisfaction of an element in a class sign of the language L – cannot be what coincides in extension with the two-place relation R_1. Nor can it even coincide with satisfaction when restricted to the subset D^* of class signs in D (though allowed to have an arbitrary extension in D otherwise). Satisfaction in this sense cannot be the interpretation of R_1 (which we are assuming to be a two-place relation), because if it were – and even without knowing any particular paradox – it would follow that the unrestricted comprehension axiom schema,

$$(\exists y)(x)(x \in y \equiv A(x))$$

holds in the domain in question. Simply take y to be the very formula $A(x_1)$ itself.[12] This would mean that the unrestricted comprehension axiom is satisfied—but it cannot be. You don't even have to know any particular paradox to obtain this conclusion.

The unsatisfiability of the unrestricted comprehension axiom schema is enough to draw Tarski's conclusion that a conventional first-order language cannot fully contain its own semantics – cannot be semantically closed.[13] It cannot contain its own satisfaction predicate, even when confined to formulae with one free variable.

Now, people didn't think of first-order structures for languages L that contained L itself as a part. We could weaken this a little bit by not necessarily assuming it literally contains the formulae of its own language L as elements – though because this is only a matter of structure, it is hard to say whether this is very important. But suppose one has a coding function, which I will suppose to be a one-to-one (though one-to-oneness isn't really necessary for everything I say) mapping of at least class signs into elements of the domain

$$f\colon A(x_1) \to D^* \subseteq D$$

(so the range is some subset D^* – it maps the class signs and maybe more (perhaps the whole language L of the structure) onto D^* – which is contained in D). As long as the domain is infinite, there will be such a function. This is purely a matter of cardinality.[14] However, if this is a countable language L, as long as D contains a countable subset, there will be plenty of such functions. Any one of them could be called a coding function. (See how we are getting towards Gödel a little bit?) Then R_1 cannot be satisfied under the coding function, but the two-place relation

$$a \text{ satisfies } f(A(x_1))$$

can't be any two-place relation definable in the language L at all – not a primitive one but not a defined one either – because then we would be able to obtain the truth of the unrestricted comprehension axiom schema. Remember, we don't need to know any particular paradox for this, just that the unrestricted comprehension axiom schema cannot be true.

So far we have assumed nothing much about D other than infinite cardinality, and it follows that either the coding function or satisfaction must be undefinable in the language L.

Tarski – I am going backwards, because Tarski was plainly inspired by (dependent on) Gödel rather than the other way around, but we are doing things in reverse order here – not only proved

the conclusion that a first-order language cannot be semantically closed (where it contains either its own syntax, so f could be the identity map, or a coding of its syntax (its formulae), so that f could be some other map – there always are such maps, due to cardinality considerations), but also drew the stronger conclusion that truth in the language was undefinable.[15] What premises do we need for that? Why should truth be undefinable? Well, satisfaction reduces to truth, given certain premises. What does it mean for a to satisfy the formula $A(x_1)$? Or putting it backwards, for $A(x_1)$ to be true of a? It simply means that if we put in a name of a in place of x_1, the resulting formula with no free variables

$$A\ulcorner a \urcorner$$

is true in L. Now, to reduce satisfaction to truth, or at least to make the reduction possible for an appropriate subset D^* of D, one including the (codes of) formulae of L (or at least the codes of the class signs), we must be able to define a naming relation in L that relates each code of a formula to a code of its name in L. As I put it this way, which is close to the way Gödel put it, we have to be able to define substitution too. Alternatively, as probably was first observed by Tarski and perhaps independently by Quine, we could define everything this way:

$$(\exists x_1)(x_1 = \ulcorner a \urcorner \wedge A(x_1))$$

(Here the function of a expressed by '$\ulcorner$' '$\urcorner$' is thought to be definable in L. In the famous Gödelian case, the natural numbers are the (codes of) formulae, and the Gödel numbers of their numerals are the codes of their names. An appropriate defining relation is available in the language.) This is an alternative way of reducing satisfaction to truth, but it still requires a naming relation definable in the language.[16] So to reduce satisfaction to truth, we need a function g mapping every class name or code of a class name into a name of the code:

$$g\colon A(x_1) \rightarrow \ulcorner A(x_1) \urcorner$$

We are well along the way now, because if truth is not definable, but provability is, under the coding (and this will certainly be true in a rich enough system, like set theory; in fact it is true in

first- order arithmetic, which is a little harder), then truth cannot be equal to provability, so they must be two different things. Hence, either there is something false but provable, which is very bad, or true but unprovable, which is also bad, but not as bad.[17] And, of course, Gödel opts for the latter case, assuming that if you have an axiomatic system in the language, all the axioms are true. The dilemma involved can be regarded as the weakest and most general form of the Gödel-Tarski result, that is, either a formal system fails to define its own basic syntactic properties, or truth fails to coincide with provability. In the standard cases the second proves to be the case, and provability is even arithmetically definable.

Given this latter fact, we can say more, even at this point. In a famous monograph, Tarski, Mostowski, and Robinson proposed a theory *R* that was supposed to be a very weak theory such that any recursively axiomatizable theory containing *R* allowed both Gödel's and Rosser's constructions to be carried out. Apparently, the real power of the theory came from the two final axiom-sets. Only one of these seemed necessary to show that the Gödel theorem held in any consistent extension; and this impression was correct for the usual form of the theorem. The last axiom-set clearly gives the impression that the authors thought it necessary to show the same thing for the Rosser form of the theorem, though plainly it was not needed for Gödel's original form of the theorem. But it turns out that the extra axiom-set was superfluous for the Rosser theorem too, since the whole theory *R* was interpretable in the theory with the last axiom-set deleted.[18]

Here I would like to introduce another system I simply call "School". It is simply theory *R* without the last two axiom sets.[19] So, it consists simply of the following tables. Assume we have 0 and a successor symbol as primitive. $0^{(m)}$ will stand for 0 followed by *m* successor symbols. There are two primitive functions, plus and times. We have as basic axioms:

$$0^{(m)} + 0^{(n)} = 0^{(m+n)}$$

$$0^{(m)} \cdot 0^{(n)} = 0^{(m \cdot n)}$$

And finally:

$$0^{(m)} \neq 0^{(n)}, \text{ if } m \neq n.$$

The point of the axioms of school is that they guarantee that any truth-function of the atomic sentences (i.e. any quantifier-free sentence of the language) is decidable in School, and in fact correctly.

Gödel showed that the provability predicates of the usual axiomatic systems were definable in terms of just plus and times. One should not regard this aspect of his work as essential to the main morals, both mathematically and philosophically, of his work, though it was certainly technically interesting. If other functions, such as exponentiation, or even other things needed to define the provability predicate of the formal system, were introduced as primitive, little would be essentially changed. And remember that his basic concern was the provability predicate for *PM*, whose definability within the system is much easier.[20]

Nevertheless, there is a point I am trying to make here. Consider any first-order system S whose axiom-set is definable in terms of plus and times, and which contains School (or really, in which School can be interpreted). It follows from any usual formulation of first-order logic that provability is also arithmetically definable. Assume S is consistent. Then by our basic result, truth in S (in the language of School plus conventional first-order logic over numbers) cannot coincide with provability in S. Hence, S either proves some false statement A, or fails to prove some true statement. But given that S is consistent, if it proves a false statement A, it fails to prove the true statement $\neg\neg A$.[21] So there is a statement that is true but unprovable in S.

Now, consider a true but unprovable statement A written in prenex form,[22] where the number of quantifiers n is a minimum for this property. n cannot be zero, since School correctly decides every quantifier-free statement. We can also show that A cannot have the form "$\exists x(Ax)$" where the initial quantifier is existential. By hypothesis the number of quantifiers in prenex form is as small as possible, for a true but unprovable statement of S. Yet, if $\exists x(Ax)$ were true, $A(0^n)$ would be true for some n. By the minimality property of the statement A as having the smallest number of quantifiers in prenex form for a true but unprovable statement, $A(0^n)$, being true, must be provable. But then $\exists x(Ax)$ follows by existential generalization, so it is provable after all, contrary to our hypothesis.

Hence, the minimal true but unprovable statement A must be of the form '$\forall x(Ax)$'. Since A is true, each instance $A(0^n)$ must be true. So, by the minimality property, each instance $A(0^n)$ must

be provable. Hence, in Tarski's terminology, *S* is ω-incomplete. It therefore is undecidable if *S* is ω-consistent.

These are the properties of Gödel's undecidable statement, obtained by a non-constructive[23] proof. Smullyan has remarked that most mathematicians have heard that Gödel found an undecidable statement, but not that he showed that there is a formula *Ax* each of whose instances is provable but not the general statement $\forall x(Ax)$ (Smullyan 1992: 73). Presumably, he meant that the mathematicians in question would find the latter result much more surprising. At least, my own personal experience with one mathematician agreed with precisely that.

Note that in this case the result cannot possibly be improved to give a Rosser form (unlike the situation with theory *R*, even with the superfluous axiom set deleted). This is true because there are complete and consistent arithmetically definable axiom systems in the language of arithmetic, but it is also true even if recursive axiomatizability is required. Notice that the axioms of School hold in the real and in the complex numbers, whose theories are complete and decidable. This was not, of course, the "intended" interpretation, but the complete and consistent non-recursive systems don't hold in the "intended" interpretation either. In the usual terminology, School, unlike *R*, is not essentially undecidable, even though it is sufficient for all its extensions to satisfy the Gödel theorem.

Up to now, we have restricted ourselves to mere knowledge that the unrestricted comprehension axiom is inconsistent, without knowing any particular paradox. But of course we do know particular paradoxes, of which Russell's paradox is the best known and simplest.

We have already remarked on the relation between satisfaction and truth, and what is needed to get a proof of the undefinability of truth from that of satisfaction. Here, however, let us consider the matter more specifically. The Grelling (heterological) paradox was of course a version of Russell's paradox.[24] Membership was replaced by satisfaction, just as it was replaced in the barber version. In this case, however, we appear to have a related semantic paradox. The following is Quine's exposition of the matter:

> To explain this paradox requires first a definition of the autological, or self-descriptive, adjective. The adjective "short" is short; the adjective "English" is English; the adjective "adjectival" is adjectival; the adjective "polysyllabic" is polysyllabic. Each of these adjectives is, in Grelling's

> terminology, autological: each is true of itself. Other adjectives are heterological; thus "long," which is not a long adjective; "German," which is not a German adjective; "monosyllabic," which is not a monosyllabic one.
>
> Grelling's paradox arises from the query: Is the adjective "heterological" an autological or a heterological one? We are as badly off here as we were with the barber. (1966: 4)

The Grelling paradox doesn't really need to apply only to single adjectives, but could apply to whole phrases, nor does it require the invention of any special word such as "heterological". As Quine points out, the paradox could be formulated in terms of "not true of self". This phrase appears to be true of itself if and only if it is not. "Not true of itself" is just "heterological" spelled out. Quine says that he considers this an antimony, that is, a real shock to our intuitions; we can't explain what is wrong with it.[25]

Now Quine turns his attention to the venerable Liar paradox. He considers various forms of it, starting with the "paradox of Epimenides the Cretan, who said that all Cretans were liars". (1966: 6). He finds possible loopholes in this formulation and tries variants such as "I am lying" and "this sentence is false". Finding possible objections to all these reformulations,[26] he finally comes up with a famous formulation of his own:

> If, however, in our perversity we are still bent on constructing a sentence that does attribute falsity unequivocally to itself, we can do so thus: "'Yields a falsehood when appended to its own quotation' yields a falsehood when appended to its own quotation". This sentence specifies a string of nine words and says of this string that if you put it down twice, with quotation marks around the first of the two occurrences, the result is false. But that result is the very sentence that is doing the telling. The sentence is true if and only if it is false, and we have our antinomy. (1966: 7)

That is a famous and very snappy form of the Liar paradox. Quine goes on to say:

> This is a genuine antinomy, on a par with the one about "heterological," or "false of self," or "not true of self," being true of itself. But whereas that earlier one turned on "true of," through the construct 'not true of self', this new one turns merely on "true," through the construct "falsehood," or "statement not true". (1966:7)

Quine's point is that it is a paradox on a par with the "heterological" one – presumably a *different* one.

But is it really different? If we have a name for an object, say "Mars," then "red" is true of Mars if and only if "Mars is red" is true, or equivalently, "is red" yields a truth after "Mars". As we have emphasized before, if we have a systematic naming function, satisfaction or ('true of') can be reduced to truth *simpliciter*. In the present case, where linguistic phrases are involved, quotation forms just such a systematic naming function. Something will be true of the phrase "is long" precisely when the corresponding sentence becomes true with "is long" filling in the blank position.

When does a phrase yield a falsehood when appended to its own quotation? Precisely when it is not true of itself. The paradoxical example really is:

> "yields a falsehood when appended to its own quotation" is not true of itself.

Here the subject phrase "yields a falsehood when appended to its own quotation" is just another way of saying "is not true of itself". This isn't a paradox *on a par* with the "heterological" paradox; it *is* the "heterological" paradox, given that 'true of' can be reduced to 'true'. Any paradox regarding membership can be changed into one of satisfaction, and then changed into one regarding truth; and the simplest one is the Russell paradox. So this is not, as it is usually thought to be, Tarski's clever and complicated form of the Epimenides paradox – it simply *is* the "heterological" paradox, using the fact that statements are nameable, here by quotation marks, and perhaps by some other form of code naming in a formal system. So, one shouldn't say it is a genuine antinomy *on a par* with the "heterological" paradox; it is the same one.[27]

Notice that Quine doesn't realize that he really has the "heterological" paradox all over again. And we don't have to think of this as a form of the Liar paradox; we don't have to notice that this is a statement that says "I am false". If someone did not notice this, there would still be a paradox, namely, the "heterological" paradox, which is what it really is, souped up. So of course it is true if and only if it is false, because all paradoxes are like that; but if someone did not notice that this was a form of the Liar paradox, we could still get the result.

Now, where, finally, does Gödel's first incompleteness theorem in the form that he stated it fit in to all this? Let us forget about basing everything on such a weak system as School. The usual presentation of Gödel's arithmetization does not rely on this. Given a suitable coding (Gödel numbering) of formulae and of proofs, we must be able to decide within the system we are studying whether a given sequence of formulae is a proof or not. The Tarski-Mostowski-Robinson system *R* is sufficient,[28] but as I have already emphasized, we should not forget that Gödel's original purpose was not to prove the incompleteness of very weak systems, even, for example, first-order arithmetic, but of stronger systems such as *PM* (see the title of his paper). Then it is much easier to see that the system can decide whether a given sequence is a proof or not. And indeed, to see the representability within the system of primitive recursive functions and predicates.

Suppose now we try to imitate the "heterological" paradox, only replacing satisfaction (or "true of") by "provability of". Replacing adjectives, or adjectival phrases, by formulae with one free variable (Gödel's "class signs" – remember that we could even fix the free variable as x_1), a class sign $A(x_1)$ is naturally called "provable of" a number n if $A(0^n)$ is provable, or alternatively, as we have seen, if $(\exists x_1)(x_1 = 0^{(n)} \wedge A(x_1))$ is provable. Suppose that we identify formulae with their Gödel numbers. Then a particular formula $Pr(x, y)$ with two free variables says that x is provable of y. $\neg Pr(x_1, x_1)$ is class sign (in Gödel's sense) that says that a formula is unprovable of itself. It itself has a particular Gödel number n, and $\neg Pr(0^n, 0^n)$ is simply a way of saying:

"Unprovable of itself" is unprovable of itself.

This is precisely the statement G constructed by Gödel.

Thus the basic statement *G* can be called Gödel's form of the "heterological" paradox, and to the present writer its content is clearer than if it is regarded in terms of the Liar paradox.[29] If our

base system, say, *PM*, is consistent, *G* cannot possibly be provable. For the proof would be a proof that "unprovable of itself" is unprovable of itself, and this very proof would show that "unprovable of itself" *is* provable of itself. Similarly, the reader can complete the argument to show that *G* cannot be disprovable either, assuming that *PM* is ω-consistent. The argument can be carried through without noticing explicitly that *G* says something about itself (i.e. that it itself is not provable).[30]

From this point of view Gödel's statement *G* is not just constructed, but makes a clear and well motivated statement. Sometimes I have used the term "Gödel heterological" for unprovable of itself, so that the statement *G* is "Gödel heterological is Gödel heterological".[31] It is a direct analogue of the "heterological" paradox, which in turn is an analogue of Russell's paradox.

Similarly, one could give an analogue of Quine's variant of Russell's paradox in terms of the class of all unreciprocated classes; one formula reciprocates a formula when it is provable of it. We have the predicate "$Pr(x, y)$" which says that x is provable of y in the sense defined above. We can call a one-place predicate (formula with one free variable or Gödelian class-sign) unreciprocated in terms of provability if there is no formula that it is provable of and of which it is provable, i.e. $\neg(\exists y)(Pr(x, y) \wedge Pr(y, x))$. Abbreviate this by $Unr(x)$, a Gödelian class-sign. Let m be the Gödel number of this class-sign and let $Unr(0^m)$ be the formula G^*. Once again, G^* makes a universal statement, each instance of which is checkable. Now can G^* be provable? Not if the underlying system is consistent. For if G^* were provable then $Unr(x)$ would by definition be provable of itself and hence reciprocated by itself, contrary to what it itself asserts, and hence contrary to a concrete and checkable instance of G^*. So if the underlying system is consistent, G^* is unprovable. However, what it says must be true. For it says that, for this particular formula, there is no formula that reciprocates it. If there were one, then there would be some formula *H* with the Gödel number n, and two proofs with Gödel numbers p_1 and p_2 that show that two formulae reciprocate each other. This would be a contradiction, since also these instances are checkable within the system. And what it implies is that *H* is unreciprocated, when in fact it also shows that G^* reciprocates it. Since G^* is a true purely universal statement, each numerical instance of which is verifiable in the system, it follows that if the system is ω-consistent G^* cannot be disprovable either.

In the case of *G*, Rosser showed that if "provable of" is replaced by the more complicated predicate "provable of with no shorter disproof of", where "shorter" is given in terms of the

Gödel numbering, the statement can be shown to be undecidable without assuming ω-consistency.[32]

References

Cobham, A. (1960). "Effectively Decidable Theories". In *Summaries of Talks Presented at the Summer Institute of Symbolic Logic,* Cornell University, 1957, 2nd edn, Princeton, Institute for Defense Analyses: 391-5.

Feferman, S., Dawson Jr., J.W., Kleene, S. C., Moore, G. H., Solovay, R. M., and van Heijenoort, J. (eds.) (1986). *Kurt Gödel. Collected Works, Volume I. Publications 1929-1936*. New York, Oxford University Press.

Frege, G. (1902). "Letter to Russell". In van Heijenoort (1967): 127-128.

Gödel, K. (1931). "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I". *Monatshefte für Mathematik und Physik* 38: 173–198. Translated in Feferman, *et al*. (1986) as "On formally undecidable propositions of *Principia Mathematica* and related systems I": 144-195.

Gödel, K. (1947). "What is Cantor's continuum Problem?" *The American Mathematical Monthly* 54: 515-25. Reprinted in Feferman, *et al.* (1986).

Hausdorff, F. (1914). *Grundzüge der Mengenlehre*. Leipzig, Von Veit.

Hausdorff, F. (1957). *Mengenlehre,* 3rd edn; translated as *Set Theory*. New York, Chelsea Pub. Co.

Kripke, S. A. (1975a). "Three Lectures on Truth". This volume chapter 9.

Kripke, S. A. (1975b). "Outline of a Theory of Truth". *Journal of Philosophy* 72(19): 690-716. Reprinted in Kripke (2011).

Kripke, S. A. (2011). *Philosophical Troubles: Collected Papers, Volume 1*. New York, Oxford University Press.

Moore, G. and Garciadiego, D. (1981). "The Burali-Forti Paradox: a reappraisal of its origins". *Historia Mathematica* 8: 319-350.

Quine, W.V.O. (1940). *Mathematical Logic*. Cambridge, MA, Harvard University Press.

Quine, W.V.O. (1962). "The Ways of Paradox". *Scientific American* 206: 84-96. Reprinted in Quine (1966).

Quine, W.V.O. (1966). *The Ways of Paradox and Other Essays*. Cambridge, MA, Harvard University Press.

Russell, B. (1902). "Letter to Frege". In van Heijenoort (1967): 124-125.

Russell, B. (1905). "On Denoting". *Mind* 14: 479–93.

Smullyan, R. (1961). *Theory of Formal Systems*. Princeton, NJ: Annals of Mathematics Studies, Princeton University Press.

Smullyan, R. (1992). *Gödel's Incompleteness Theorems*. New York, Oxford University Press.

Tarski, A. (1935). "Der Wahrheitsbegriff in den formalisierten Sprachen". *Studia Philosophica* 1: 261–405. Translated as "The Concept of Truth in Formalized Languages" by J.H. Woodger in Tarski (1983): 152–278.

Tarski, A. (1944). "The Semantic Conception of Truth and the Foundations of Semantics". *Philosophy and Phenomenological Research* 4: 341-376.

Tarski, A. (1983). *Logic, Semantics, Metamathematics*. Second edition, Corcoran, J. (ed.), Indianapolis, Hackett.

Tarski, A., Mostowski, A., and Robinson, R. M. (1953). *Undecidable Theories.* Amsterdam, North Holland.

van Heijenoort, J. (1967). *From Frege to Gödel: Source Book in Mathematical Logic, 1879–1931*. Cambridge, MA, Harvard Univ. Press.

Wang, H. (1987). *Reflections on Kurt Gödel*. Cambridge, MA, MIT Press.

Whitehead, A. N., and B. Russell. (1910, 1912, 1913). *Principia Mathematica, 3 Volumes*. Cambridge: Cambridge University Press. Second edition: 1925 (Vol. 1), 1927 (Vols. 2 and 3).

---

* This is a post-peer-review, pre-copyedit version of a chapter first published in *Naming, Necessity and More. Explorations in the Philosophical Work of Saul Kripke*, Berg, J. (ed.), Palgrave Macmillan, London, 2014: 223-241, doi: 10.1057/9781137400932_11. The definitive publisher-authenticated version is available online at: https://doi.org/10.1057/9781137400932_11.

The present paper is loosely based on a transcript of a talk delivered at the Hebrew University, in Jerusalem, Israel, on June 18, 2006. I have not entirely tried to abandon the conversational tone of the original. A version of this lecture was first given at the conference "Naming, Necessity and More" held at the University of Haifa, Israel, on June 21-24, 1999, and was subsequently given to other audiences.

[1] Many now wish to replace the old term 'recursive' with 'computable'. Though I have done this sometimes, I am too used to the old terminology to make the change consistently.

[2] As everyone knows, Gödel eventually codes finite sequences of natural numbers into single numbers. It is interesting that the coding system he uses in the preliminary sketch is not the famous Gödel numbering used in the detailed version that follows, but it is closer to the types of coding used by Quine (1940) and Smullyan (1961 and 1992). However, to see this, some modification of the way Gödel presents it would be required. Personally, I prefer the type of Gödel numbering introduced by Quine and Smullyan, and have my own variant of it (see Smullyan 1992: 46). I would also note that for most systems (those containing a Tarskian weak second-order logic), coding finite sequences of numbers into single ones is not really necessary, though it has become famous. Nor need one code finite sequences of such sequences. Gödel's basic construction could still remain.

[3] Notice that Gödel here and elsewhere in the paper uses a horizontal sign above a formula to indicate negation. His official notation, however, see p. 150, is the usual '~', and it is used when the system *P* (his version of *Principia*) is officially set up (e.g. pp. 155-7). On the other hand, on pp. 159 and 161 Gödel evidently uses '~' for the material biconditional, and on p. 161 assumes that negation is expressed by a raised horizontal line. This is a bit confusing in a paper that otherwise is set out with great precision. My failure to notice this led me to use a raised horizontal line in another sense, but I hope I have fixed this. If not, let the context show what I mean.

[4] Actually, I disagree with the suggestion made by Gödel's footnote (that is, that a statement saying directly of itself that it is unprovable would involve a faulty circularity). I think we can carry out his construction using a formula directly saying of itself that it is unprovable. See Kripke (1975b: 693, note 6). Some people have advised me that my note deserves elaboration, as perhaps I will do elsewhere. Some such elaboration is in Kripke (1975a).

One version of the construction mentioned, using a non-standard Gödel numbering, is independently due to Smullyan. In that footnote (1975b), I do say that I don't maintain that a proposition (as opposed to a sentence, a linguistic object) can be directly self-referential in this sense, and perhaps this is what Gödel may have partly in mind.

[5] In note 14 Gödel says that "every epistemological antinomy can likewise be used for a similar undecidability proof". I take "epistemological antinomy" to mean the same as what we would now call "semantic paradox".

[6] The Burali-Forti paradox is a special case, since in the original theory there is no special reduction of ordinals to sets (or classes) of a certain kind. Moreover, there is the question of reducing relations to sets. There are various possible theories, including no reductions at all. In the latter case, the Burali-Forti paradox would not simply be a consequence of the unrestricted comprehension axiom schema. For the history of this paradox, see Moore and Garciadiego (1981). (Although I agree with them that Cantor did not regard the Burali-Forti result – even with Burali-Forti's erroneous definition of a well ordered set corrected – as paradoxical, I do think he was aware of the mathematical fact that the ordinals cannot themselves form yet another well ordered set).

[7] Hausdorff refers to the "so called paradoxes of set theory" in Hausdorff (1914; 1957: 6, 29-30).

[8] Gödel's attitude was that the paradoxes "are a very serious problem, but not for Cantor's set theory" (Gödel 1947: 518). I agree (though with some reservations about the usual formulation of set theory).

[9] No relation to Gödel's use of *K* above.

[10] For a long time I had never heard of insurance companies that insured other insurance companies against catastrophic claims. (They do exist.) Russell's paradox shows that there couldn't be an insurance

company ("Russell's Insurance Company") that insures all and only those insurance companies that do not insure themselves.

[11] Russell, of course, showed by his theory of descriptions that function letters and constants could be eliminated. See Russell (1905) and Whitehead and Russell (1910/1925, vol. 1: 30-2, 66-71, 173-5).

[12] I have not bothered putting the subscript on the variable in the unrestricted comprehension axiom schema. It is clear that it would hold, regardless of subscript.

[13] The matter is often stated in the form that semantically closed languages are inconsistent. But it is really axiomatic formal theories, not interpreted first-order languages that are not consistent. My own way of putting the matter would be that semantically closed languages, formulated in the usual first-order logic, do not exist. (I make this point in Kripke (1975a).)

[14] Put this way the result depends on the axiom of choice. Otherwise, we assume that D is Dedekind infinite, or, equivalently, has a countable subset. A weak use of choice is needed to establish the equivalence of this property with the infinity of D.

[15] Actually, I am probably being too conventional in attributing this result to Tarski. According to Wang (1987: 90), who himself refers to a letter from Gödel to Zermelo (reproduced with comments by Grattan-Guinness), Gödel actually showed that a usual language cannot define its own truth in 1930, before going on to the incompleteness theorem.

[16] We don't need names in the literal sense; in Russell's theory of descriptions predicates simply uniquely satisfied by the object, for example, could take the place of names. (What we would need is a binary relation $R(x, y)$, interpreted as saying that x uniquely satisfies the predicate y. So we need a variable y ranging over one-place predicates in the language.)

[17] But see also the remark on this dilemma below.

[18] The theory R in the classical monograph of Tarski et al. (1953) consisted of the axioms of School as presented in my text together with two axiom-sets:

$$\Omega_4 \quad x \leq 0^{(n)} \supset x = 0 \text{ v } x = 0' \text{ v } \ldots x = 0^n$$
$$\Omega_5 \quad x \leq 0^{(n)} \text{ v } 0^n \leq x$$

Here the variable x is thought of as universally quantified and for any terms t_1 and t_2, $t_1 \leq t_2$ abbreviates $(\exists w)\ (w + t_1 = t_2)$, where w is a variable not occurring in t_1 and t_2.

The first four axioms including Ω_4 are needed to show that Gödel's original form of his theorem is provable for any system in which R is interpretable. Plainly, the authors thought that Ω_5 was needed to obtain the Rosser form, and hence that R is essentially undecidable. However, Alan Cobham (1960) showed that R was interpretable in R- (the system with Ω_5 omitted), and hence that we have nothing here that differentiates between the Gödel and the Rosser forms. The system School in the text does differentiate between the two, for one version of the Gödel theorem.

[19] Though I only later on noticed the relation with R.

[20] Suppose Gödel's first incompleteness theorem had been proved only for first-order arithmetic. This might have been ascribed to a weakness in the system. Stronger systems, such as *PM*, still might decide all relevant statements.

[21] This observation is due to D. A. Martin, and replaced a more complicated argument used before.

[22] That is, a bunch of quantifiers followed by something with no quantifiers: $(Q_1x_1) \ldots (Q_nx_n)(\ldots)$

[23] That this form of the result needs to be non-constructive (i.e. that the independent statement cannot be found by recursive function, given, say, an arithmetical index for the axioms of *S*) was shown, when it came up in a version of this lecture, by D. A. Martin. His proof uses the recursion theorem. The present writer found another proof, perhaps more conceptual, but I recall somewhat longer.

[24] I believe that Grelling's paradox originated in his thesis. The paradox is normally stated for natural language (see below), but could just as well be stated for a formal language. The same type of transformation could be applied to other instances of the unrestricted comprehension axiom schema.

[25] Similarly to the relation of the Russell paradox to the Grelling "heterological" paradox, we could define a semantic analogue of the "unreciprocated" paradox. An adjective, or adjectival phrase, could be called "reciprocated" if it is true of an adjective (or phrase) that is true of it; otherwise, "unreciprocated". Then, is "unreciprocated" unreciprocated? The question leads to a contradiction.

[26] I am not certain that Quine really finds all these ways out completely cogent.

[27] In some sense, Tarski seems to have realized the connection. In Tarski (1935: 248, note 2), pretty clearly an addition to the original paper since it refers to a specific construction in a paper written later (Tarski 1944), Tarski seems to see that, in natural language, something analogous to the usual proof of the impossibility of a language containing its own truth definition really uses the Grelling paradox. He does not appear to apply his observation to formal languages, however. (I owe this remark to Joseph Almog.)

[28] The main point of the system is of course that it gives an axiom showing that bounded quantifiers are in fact eliminable in favor of finite disjunctions, and hence statements all of whose quantifiers are bounded are decidable.

[29] As is usual, given what Gödel says. The same goes for the Richard paradox.

[30] What is really used is that the statement is Π_1^0, or more loosely for present purposes, that it is a universal statement each instance of which can be checked within the system. In particular, the argument uses the fact that a proof of *G* can be recognized within the system.

[31] Nathan Salmon has suggested "hetero-Gödel".

[32] In 2005 I gave a version of this lecture in Buenos Aires. Alberto Moretti wondered whether it would be necessary to produce a particular undecidable statement as Gödel did given the non-constructive proof of Gödel's theorem, assuming that one was not at all worried about constructive proofs. However, not only is it nice to have a particular example, but it is needed for Gödel's second incompleteness theorem. For one proves within *PM* (say) $Con(PM) \rightarrow G$. Hence, if *PM* is consistent, *G* is unprovable, and hence so is $Con(PM)$. For this one needs a particular statement *G* for *PM*, or whatever system is in question. (Nor is the non-constructive proof obviously formalizable in the system as stated, since it uses the notion of arithmetical truth or definability, which in a weak enough language is not itself definable. It is nice of course that Gödel obtained a Π_1^0 statement and this is essential to the second incompleteness theorem as stated.)

I would like to thank Jonathan Berg for transcribing the original lecture. My thanks to Gary Ostertag and especially to Romina Padró for their help in producing the present version. This paper has been completed with support from the Saul A. Kripke Center at The City University of New York, Graduate Center.