

Diagonalization from Cantor to Gödel

Haim Gaifman

There is a direct route from Cantor's original diagonal method to Gödel's construction that underlies the proof of the incompleteness theorem. Gödel's original paper gives an indication of it in the introductory section.¹ Here is a fuller description.

Consider an infinite sequence of classes of natural numbers, indexed by a subscript that runs over the natural numbers:

$$X_0, X_1, \dots, X_n, \dots$$

Let us write ' $X(y)$ ' for ' $y \in X$ '. Cantor considers the class X^* of all n 's such that $n \notin X_n$, in other words, all n 's for which *not* $X_n(n)$. Suppose that, for some k , $X^* = X_k$; then we get a contradiction by asking whether $X_k(k)$:

$$X_k(k) \Leftrightarrow k \text{ has the property defining } X^* \Leftrightarrow \text{not } X_k(k)$$

This shows that X^* is not in the sequence of classes.

X^* is defined by *diagonalization*, that is, through a condition that involves the predication of ' $X_n(\)$ ' of its own representative number, n . The last step in the argument employs diagonalization as well: we predicate ' $X_k(\)$ ' of k .

	0,	1,	2,	...,	n,	...
X_0	$X_0(\mathbf{0})$	$X_0(\mathbf{1})$	$X_0(\mathbf{2})$		$X_0(n)$	
X_1	$X_1(\mathbf{0})$	$X_1(\mathbf{1})$	$X_1(\mathbf{2})$		$X_1(n)$	
X_2	$X_2(\mathbf{0})$	$X_2(\mathbf{1})$	$X_2(\mathbf{2})$		$X_2(n)$	
\vdots				\ddots		
X_n	$X_n(\mathbf{0})$	$X_n(\mathbf{1})$	$X_n(\mathbf{2})$		$X_n(n)$	
\vdots						\ddots

Now suppose that the classes are those defined by the wffs of a formal language, which underlies some deductive system. Arrange in a sequence all wffs with one free variable:

$$f_0(x), f_1(x), \dots, f_n(x), \dots$$

The diagonalization now involves the substitution of the free variable by the (name in the language of) the subscript of the wff:

¹ A move of this kind, tracing the diagonal argument from Cantor's construction, through Grelling's paradox to the Liar is described in my 1983 *Erkenntnis* paper "Paradoxes of Infinity and Self-Application" (vol. 20, pp. 131-155).

$$\mathbf{f}_n(\underline{n}),$$

where \underline{n} is the name (in the formal language) of n .

Since we are speaking about wffs, rather than directly about classes and numbers, we have to rephrase ' $X_n(m)$ ' as: ' m is in the class defined by \mathbf{f}_n ', or as ' \mathbf{f}_n is true of n ', or ' $\mathbf{f}_n(\underline{n})$ is true'. The class X^* of Cantor's proof is defined by:

$$(I) \quad X^*(n) \Leftrightarrow \mathbf{f}_n(\underline{n}) \text{ is not true.}$$

By Cantor's diagonal argument, X^* is not in the sequence, i.e., it is not definable by any wff of the language. This implies also that the class of all n 's such that $\mathbf{f}_n(\underline{n})$ is true is undefinable by any wff, else the negation of this wff would define X^* . Therefore, there is no wff $\psi(x)$, such that, for all n , $\psi(\underline{n})$ says that $\mathbf{f}_n(\underline{n})$ is true.

Nonetheless, working inside the formal system, we can describe the sentences $\mathbf{f}_n(\underline{n})$ as syntactic items, which depend on n , where n appears as a variable. We can, moreover, speak inside the system about their provability, unprovability, or any other property expressible in the language. (This requires that the formal system be sufficiently expressive, which, it turns out, is not a very strong condition.) Trying to imitate Cantor's construction we substitute 'not true' by 'not provable'. We then get a wff $\psi(x)$ such that, for all n , $\psi(\underline{n})$ says that $\mathbf{f}_n(\underline{n})$ is not provable. This, more concretely, means that for each n the following is provable (and we can construct the proof effectively):

$$(II) \quad \psi(\underline{n}) \Leftrightarrow \text{'}\mathbf{f}_n(\underline{n})\text{' is not provable.}$$

Here ' $\mathbf{f}_n(\underline{n})$ ' is the name in the system of the sentence $\mathbf{f}_n(\underline{n})$. Since ψ is a wff of the language, $\psi = \mathbf{f}_k$, for some k . Thus, the following is provable for each n :

$$(II') \quad \mathbf{f}_k(\underline{n}) \Leftrightarrow \text{'}\mathbf{f}_n(\underline{n})\text{' is not provable.}$$

And, for $n=k$:

$$(III) \quad \mathbf{f}_k(\underline{k}) \Leftrightarrow \text{'}\mathbf{f}_k(\underline{k})\text{' is not provable.}$$

We end up with a sentence, namely $\mathbf{f}_k(\underline{k})$, which says of itself that it is not provable.

The argument now proceeds in the familiar way: If $\mathbf{f}_k(\underline{k})$ is provable, there is a sequence of wffs that constitutes its proof; we can prove in the system that this sequence is a proof of $\mathbf{f}_k(\underline{k})$, hence we can prove: ' $\mathbf{f}_k(\underline{k})$ is provable'. But the proof of $\mathbf{f}_k(\underline{k})$, together with the proof of (III), yields a proof of: ' $\mathbf{f}_k(\underline{k})$ is not provable'. Which implies that the system is not consistent. On the other hand, if the negated sentence, $\neg\mathbf{f}_k(\underline{k})$, is provable, then, assuming that the system is consistent, there is no proof of $\mathbf{f}_k(\underline{k})$; for each sequence of wffs the sentence saying that it is not a proof of $\mathbf{f}_k(\underline{k})$ is provable. On the other hand, the sentence saying that there is a proof is provable (since it is derivable from $\neg\mathbf{f}_k(\underline{k})$). Which means that the system is ω -inconsistent.

Now, there is no need to consider a special enumeration of the wffs. Working inside the system, the index, n , of the wff $\mathbf{f}_n(x)$ can be replaced by the wff itself. All that we need is a wff, $\psi(x)$, which says the following:

The sentence obtained from the wff x , by substituting the free variable of x by the name of x , is not provable.

If $\ulcorner \mathbf{f} \urcorner$ is the name of \mathbf{f} in the formal language, then, in a more formal notation, (II) becomes the following biconditional:

$$(II^*) \quad \psi(\ulcorner \mathbf{f} \urcorner) \leftrightarrow \neg \text{Provable}(\ulcorner \mathbf{f} (\ulcorner \mathbf{f} \urcorner) \urcorner)$$

This biconditional is provable for each wff $\mathbf{f}(x)$. Choosing $\mathbf{f} = \psi$, the following is provable:

$$(III^*) \quad \psi(\ulcorner \psi \urcorner) \leftrightarrow \neg \text{Provable}(\ulcorner \psi (\ulcorner \psi \urcorner) \urcorner)$$

Hence $\psi(\ulcorner \psi \urcorner)$ is a sentence saying of itself that it is not provable; if $\mathbf{g} = \psi(\ulcorner \psi \urcorner)$, then $\mathbf{g} \leftrightarrow \neg \text{Provable}(\ulcorner \mathbf{g} \urcorner)$ is provable.

If the language provides for making terms of the form: *the unique u such that $\alpha(u)$* , then we can form a term, say $\text{diag}(x)$, which reads:

The sentence obtained from the wff x by substituting the free variable of x by the name of x .

And then $\psi(x) = \neg \text{Provable}(\text{diag}(x))$. But the use of such a term is not essential. We can express the same thing, provided that we have a wff, say $\delta(x,y)$, which formalizes:

y is the sentence obtained from the wff x by substituting the free variable of x by the name of x .

In that case $\psi(x) = \forall y [\delta(x,y) \rightarrow \neg \text{Provable}(y)]$. The sentence \mathbf{g} is then obtained by substituting the variable x in $\psi(x)$ by $\ulcorner \psi(x) \urcorner$.

The same construction applies if, instead of $\neg \text{Provable}(x)$, we use any wff $\alpha(x)$ of the language. We construct a wff $\beta(x)$ such that, for every $\mathbf{f}(x)$, the following biconditional is provable:

$$(II^{**}) \quad \beta(\ulcorner \mathbf{f} \urcorner) \leftrightarrow \neg \alpha(\ulcorner \mathbf{f} (\ulcorner \mathbf{f} \urcorner) \urcorner)$$

If we can use the term $\text{diag}(x)$, we can put: $\beta(x) = \alpha(\text{diag}(x))$. Otherwise,

$$\beta(x) = \forall y [\delta(x,y) \rightarrow \alpha(y)].$$

Choosing in (II**) $f = \beta$, we get the following provable biconditional:

$$(III**) \beta(\ulcorner \beta \urcorner) \leftrightarrow \neg \alpha(\ulcorner \beta(\ulcorner \beta \urcorner) \urcorner)$$

Hence $\beta(\ulcorner \beta \urcorner)$ says of itself that it is “an α ”.

The usual presentation of Gödel’s results goes the other way round. One starts by stating the goal of constructing a sentence that says of itself that it is unprovable (i.e., for which $g \leftrightarrow \neg \text{Provable}(\ulcorner g \urcorner)$ is provable). This goal is achieved by proving a general result, the so-called diagonalization lemma, which shows how to get, given any $\alpha(x)$, a sentence σ such that $\sigma \leftrightarrow \alpha(\ulcorner \sigma \urcorner)$ is provable. But then the construction seems a magic trick.

The above constructions have (as Quine pointed out, for the case of ‘untrue’) natural language analogues. Observe that in many natural languages the name of a linguistic expression is obtained by putting it in quotes, and that predication in English is effected by putting the phrase that does the predication after the noun (or noun phrase), as in ‘Snow is white’. Therefore, let us define the result of *applying an expression to itself* as the quoted expression followed by the expression itself. For example:

- Applying ‘runs’ to itself yields: ‘ ‘runs’ runs’.
- Applying ‘is white’ to itself yields: ‘ ‘is white’ is white’.
- Applying ‘is an English phrase’ to itself yields: ‘ ‘is an English phrase’ is an English phrase’.
- Applying ‘boy’ to itself yields ‘ ‘boy’ boy’.

The first three cases yield English sentences; the third sentence is true; the other two (under the usual reading of ‘runs’ and ‘white’) are either false or category mistakes; the fourth is gibberish.

Now consider the following expression, where ‘ Φ ’ stands for some unspecified description,

yields when applied to itself a Φ

It expresses a certain property of English expressions: the property of yielding a Φ when applied to itself. We can now form a sentence stating that the above expression has itself this property:

‘yields when applied to itself a Φ ’ yields when applied to itself a Φ .

Let ‘ Γ ’ stand for this sentence. Γ states that Δ is a Φ , where Δ is the expression obtained by applying to itself the expression ‘yields when applied to itself a Φ ’. To find Δ , we apply this expression to itself; according to the prescription, we get:

‘yields when applied to itself a Φ ’ yields when applied to itself a Φ .

But this is no other than Γ ! Therefore, we have just shown:

Γ if and only if ‘ Γ ’ is a Φ .

Thus, we have constructed a sentence that says of itself that it is a Φ . Now, in natural language we can construct self-referential sentences by writing, or “putting” sentences in locations; for example, by putting in location L the sentence: ‘The sentence in location L is a Φ ’. But this possibility is, in general, not available in formal systems. The diagonalization technique enables us to achieve the same effect, both in formal and natural languages, without pointing to sentences through locations, time points, speakers, or what have you.

Note that, in natural language, the status of the sentence Γ depends crucially on Φ . If ‘ Φ ’ stands for ‘non-true sentence’, then Γ is provably equivalent to: ‘ \mathbf{G} ’ *is not true*; we therefore are led to the Liar paradox. If ‘ Φ ’ stands for ‘true sentence’, the resulting Γ does not lead to contradiction, but many philosophers would argue that it expresses no true-or-false proposition. For these reasons, in formal consistent systems ‘true’ cannot be available as a truth predicate for the entire language (if “truth predicate” is supposed to imply certain standard features). ‘Provable’ *is* available in formal systems, because the notion of formal proof is completely and rigorously specified. Not so in natural language, where it is open ended and can give rise to the same kind of trouble that are generated by ‘true’. For example, one tend to take for granted that a logical contradiction is not provable; because anything that “proves” a contradiction cannot, by virtue of this very fact, be considered a proof. But then, take the sentence, \mathbf{g} , which says of itself that it is not provable. The argument above shows that if it is provable, we can prove contradiction. Since the latter is impossible, the sentence is indeed not provable. But this is equivalent to the sentence itself. Therefore we have just proved \mathbf{g} , and, consequently we get a contradiction.

On the other hand ‘ Φ ’ can stand for numerous descriptions without giving rise to particular difficulties. The truth-value of the resulting self-referring sentences is non-problematic. Consider for example the following:

1. ‘yields when applied to itself a sentence containing twenty words’ yields when applied to itself a sentence containing twenty words.
2. ‘yields when applied to itself a sentence that contains twenty words’ yields when applied to itself a sentence that contains twenty words.
3. ‘yields when applied to itself a sentence containing two times ten words’ yields when applied to itself a sentence containing two times ten words.
4. ‘yields when applied to itself a puzzling sentence’ yields when applied to itself a puzzling sentence.
5. ‘yields when applied to itself a boring sentence’ yields when applied to itself a boring sentence.

1. is true (check by counting the words), 2. and 3. are false. Any unclarity concerning the truth values of 4. and 5. is due to the vagueness, or ambiguity of ‘puzzling’ and ‘boring’, not to the self-referential nature of the sentences.