Gödel's Incompleteness Theorem

Lecture notes for COM3412 *Logic and Computation*

4th May 2004

1 What did Gödel prove?

Gödel proved that first-order arithmetic[1] cannot be completely axiomatised. More exactly, no finitely-specifiable axiom system for first-order arithmetic can be both sound and complete.

Note that if we had a sound and complete axiomatisation for first-order arithmetic, then we could mechanically churn out all the first-order truths of arithmetic, which would therefore be at least a semi-decidable theory. The set of true formulae of first-order arithmetic would be recursively enumerable. Gödel's result is thus tantamount to showing that the set of truths of first-order arithmetic is not recursively enumerable. This means that there is no effective procedure for generating all and only the true formulae of first-order arithmetic. In effect, it means that arithmetic cannot be completely mechanised; there is always room for ingenuity or creativity in devising new methods of proof.

2 How did he prove it?

Suppose that we have a formal system S which is rich enough to be used to state and prove formulae of the first-order arithmetic of the natural numbers. We must show that S cannot be both sound and complete.

- Each formula ϕ in S has a *standard interpretation* which is a proposition p_ϕ of arithmetic.
- We write $S \vdash \phi$ to mean that S provides a proof for ϕ.
- We say S is *sound* if every formula it provides a proof for is true under the standard interpretation (i.e., $S \vdash \phi$ implies p_ϕ).
- We say S is *complete* if S provides a proof for every formula which is true under the standard interpretation (i.e., p_ϕ implies $S \vdash \phi$).

In addition to the standard interpretation, Gödel showed how to construct, for certain formulae ϕ, an alternative interpretation, which is a statement q_ϕ *about the system* S, such that q_ϕ is true if and only if p_ϕ is true. We shall call this alternative interpretation the *Gödelian interpretation*.

Gödel's achievement was to set up the Gödelian interpretation in such a way that there is a formula γ in S whose Gödelian interpretation q_γ is the statement "γ is not provable in S". The formula γ is known as the Gödel formula for the system S.[2] Like all formulae in S, it also has a standard interpretation p_γ, which is an arithmetical statement which may or may not be true. Suppose p_γ is true. Then q_γ must

[1] I.e., the system described in §5 of *Some Computational Aspects of Logic*.

[2] A system does not have a unique Gödel formula: which formulae of a system can function as Gödel formulae for that system depends sensitively on the way in which the Gödelian interpretation is defined—for more details, see below. What is important, though, is that one can always construct a Gödel formula for any system powerful enough to express and prove statements of first-order arithmetic.

be true also, which means that γ is not provable in S. Hence there is a formula that is true under the standard interpretation but which is not provable in S, which means that S is not complete. Suppose on the other hand that p_γ is false. Then q_γ must be false also, which means that γ is provable in S. Hence there is a formula that is provable in S but which is not true under the standard interpretation, which means that S is not sound. Since p_γ must be either true or false, it follows that S is either incomplete or unsound. In any case, it cannot be both sound and complete.

3 What is the Gödelian interpretation?

To quote from Gödel himself[3],

> The formulas of a formal system (we restrict ourselves here to the system PM) in outward appearance are finite sequences of primitive signs ..., and it is easy to state with complete precision *which* sequences of primitive signs are meaningful formulas and which are not. Similarly, proofs, from a formal point of view, are nothing but finite sequences of formulas (with certain specifiable properties). Of course, for metamathematical considerations 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. The metamathematical notions (propositions) thus become notions (propositions) about natural numbers or sequences of them; therefore they can (at least in part) be expressed by the symbols of the system PM itself. In particular, it can be shown that the notions "formula", "proof array", and "provable formula" can be defined in the system PM; that is, we can, for example, find a formula $F(v)$ of PM such that $F(v)$, interpreted according to the meaning of the terms of PM, says: v is a provable formula. We now construct an undecidable proposition of the system PM, that is, a proposition A for which neither A nor not-A is provable, in the following manner.
>
> A formula of PM with exactly one free variable, that variable being of the type of the natural numbers ..., will be called a *class sign*. We assume that the class signs have been arranged in a sequence in some way, we denote the nth one by $R(n)$, and we observe that the notion "class sign", as well as the ordering relation R, can be defined in the system PM. Let α be any class sign; by $[\alpha; n]$ we denote the formula that results from the class sign α when the free variable is replaced by the sign denoting the natural number n. The ternary relation $x = [y; z]$, too, is seen to be definable in PM. We now define a class K of natural numbers in the following way:
>
> $$n \in K \equiv \overline{Bew}[R(n); n] \tag{1}$$
>
> (where $Bew\ x$ means: x is a provable formula)[4]. Since the notions that occur in the definiens can all be defined in PM, so can the notion K formed from them; that is, there is a class sign S such that the formula $[S; n]$, interpreted according to the meaning of the terms of PM, states that the natural number n belongs to K. Since S is a class sign, it is identical with some $R(q)$; that is, we have
>
> $$S = R(q)$$
>
> for a certain natural number q. We now show that the proposition $[R(q); q]$ is indecidable in PM. For let us suppose that the proposition $[R(q); q]$ were provable; then it would also be true. But in that case, according to the definitions given above, q would belong to K, that is, by (1), $\overline{Bew}[R(q); q]$ would hold, which contradicts the assumption. If, on the other hand, the negation of $[R(q); q]$ were provable, then $\overline{q \in K}$, that is, $Bew[R(q); q]$, would hold. But then $[R(q); q]$, as well as its negation, would be provable, which again is impossible.

This occurs in §1 of the paper; §2 begins "We now proceed to carry out with full precision the proof sketched above".

[3]This is from the seminal paper, published in 1931, in which Gödel announced his result, "*On Formally Undecidable Propositions of* Principia Mathematica *and Related Systems*". It is reprinted in Jean van Heijenoort (ed.), *From Frege to Gödel, a Source Book in Mathematical Logic, 1879–1931*, Harvard University Press, 1967, and also in M. Davis, *The Undecidable*, Raven Press, New York, 1965, and in S. G. Shanker, *Gödel's Theorem in Focus*, Routledge 1989.

[4]The bar denotes negation.

A couple of observations about the above passage.

- PM is the system presented in the *Principia Mathematica* of Bertrand Russell and A. N. Whitehead (1910), a monumental, but ultimately unsuccessful, attempt to derive the whole of mathematics from pure logic. Although Gödel's explicit construction works with this system, he shows that the same thing can be done with *any* system powerful enough to express the first-order truths of arithmetic. (This does not include pure first-order logic, which Gödel himself proved was complete; but it does include pure second-order logic, in which the arithmetical notions of "natural number", "successor", "addition" and "multiplication" can be defined.)
- Why Bew? Gödel was writing in German, in which the word for "provable" is *Beweisbar*.

4 Gödel Numbering

As the passage above explains, Gödel defined the alternative, "Gödelian" interpretation of arithmetical expressions by *encoding* all such expressions as numbers. In that way, a formula which ascribes a certain arithmetical property to some number might also, in certain cases, be interpretable as a ascribing some logical property to the formula encoded by that number. Gödel's scheme of encoding relies on what is known as the *Fundamental Theorem of Arithmetic*, which states that every positive integer can be expressed as the product of primes in one and only one way (e.g., $1446480 = 2^4 \times 3^2 \times 5 \times 7^2 \times 41$). Gödel first assigned an odd number to each of the primitive symbols of the system. To illustrate, we shall use a somewhat different notation from Gödel's; our language will have thirteen primitive symbols to which we assign odd numbers as follows:

0	s	$=$	$\neg$	$\vee$	$\forall$	$($	$)$	$+$	$\times$	x	$'$
1	3	5	7	9	11	13	15	17	19	21	23

Here s is for suc, and we can form an unlimited number of variables as $x, x', x'', x''', \ldots$. The symbols "$\wedge$", "$\rightarrow$", "$\leftrightarrow$", and "$\exists$" can be defined in terms of "$\neg$", "$\vee$", and "$\forall$". In this pared-down language, the formula $\exists x(x + x = suc(x))$ would be expressed as "$\neg\forall x\neg x + x = sx$". The number assigned to this formula is computed by taking the first prime, 2, and raising it to the power given by the first symbol in the expression, multiplying by the second prime, 3, raised to the power corresponding to the second symbol, and so on, giving us the (enormous!) number

$$2^7 \times 3^{11} \times 5^{21} \times 7^7 \times 11^{21} \times 13^{17} \times 17^{21} \times 19^5 \times 23^3 \times 29^{21}.$$

Since this number can *only* be factorised in this way, one can uniquely retrieve the formula if given the number. We call it the **Gödel number** of the formula. For a formula ϕ, we shall write $g(\phi)$ to denote its Gödel number. Note that not all numbers are the Gödel numbers of formulae. For example, the Gödel number of any formula must be even: this is because it is divisible by 2^n, where n is the number assigned to the first symbol in the formula.

Gödel next assigns numbers to *sequences* of formulae (such as one finds, for example, in proofs). The Gödel number assigned to a sequence $\phi_1, \phi_2, \phi_3, \ldots, \phi_n$ of formulae, is

$$2^{g(\phi_1)} \times 3^{g(\phi_2)} \times \cdots \times p_n^{g(\phi_n)}.$$

Notice that since each of the the $g(\phi_i)$ is even, this number is a perfect square, whereas the Gödel number of a formula is never a square (since each of its prime factors occurs with odd multiplicity). Thus formulae can always be distinguished from sequences of formulae by their Gödel numbers.

Moreover, some *logical* properties of a formula or sequence of formulae in the system S correspond to *arithmetical* properties of its Gödel number. For example, a formula is a universal generalisation if and only if its Gödel number is divisible by 2^{11} but by no higher power of 2. Gödel's *tour de force* was to show how even quite complex logical properties such as

"x is a well-formed formula",
"x is a substitution instance of one of the axioms of S",
"x is a sequence of formulae which constitutes a correct proof in S",
"formula x is provable from the axioms of S",

correspond precisely, under his encoding scheme, to arithmetical properties of the Gödel numbers, properties which, moreover, can be expressed in the formal language of S. In this way, the logical system S, which was set up in order to talk about arithmetic, is co-opted into talking about itself, and it is because of this reflexiveness that it is possible to construct a formula which cannot be either proved or disproved within S.

5 Construction of the Gödel Formula

Here, in outline, is what Gödel did. He first showed how to construct a formula $Pf(x, y)$ (with free variables x and y), which says that (i) x is the Gödel number of a formula $\phi(z)$ containing one free variable, and (ii) y is the Gödel number of a proof of the formula $\phi(x)$ obtained by substituting for the free variable z in $\phi(z)$ the Gödel number x of that very formula $\phi(z)$. (A proof of a formula ϕ consists of a sequence of formulae, each of which is either an axiom of S or is derived from one or more earlier formulae in the sequence using a rule of inference of S, and such that the last formula in the sequence is ϕ itself.)

The next step is to consider the formula $\forall y \neg Pf(x, y)$. Note that this formula has one free variable x; let us abbreviate it $\beta(x)$, and let its Gödel number be g. The formula $\beta(x)$ says that for any y, it is not the case that y is the Gödel number of a proof in S of the formula $\phi(x)$, where x is the Gödel number of the open formula $\phi(z)$: in short, the formula $\phi(x)$ cannot be proved in the system.

Consider now the formula $\beta(g)$, i.e., $\forall y \neg Pf(g, y)$, which is obtained by substituting for x in the formula $\beta(x)$ the Gödel number of that very formula (i.e., the symbolic representation of that number in the formal language). This formula says

(1) For each y, it is not the case that g is the Gödel number of a formula $\phi(x)$ containing one free variable, such that y is the Gödel number of a proof, in S, of the formula $\phi(g)$.

Now we know that g is the Gödel number of $\beta(x)$, so we can simplify our statement (1) of what $\beta(g)$ says as follows:

(1') For each y, it is not the case that y is the Gödel number of a proof, in S, of the formula $\beta(g)$.

More simply still, $\beta(g)$ says that

(1") $\beta(g)$ cannot be proved in S.

Thus the formula $\beta(g)$ asserts its own unprovability in S; it is therefore a Gödel formula for the formal system S.

Having constructed the Gödel formula for S, we can then argue exactly as in §2 above to establish that S cannot be both sound and complete.

6 Digression: Penrose's attack on 'strong' AI.

Gödel's theorem says that if we are given a sound formal proof system for arithmetic, then we can construct a statement of arithmetic (the Gödel formula of the system) which, though true, is not a theorem of the system.

Penrose deduces from this that

> "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth." (*Shadows of the Mind*, §2.5)

More exactly, we can prove the following theorem:

Theorem. *It is not the case that I can correctly know, of some sound formal proof system whose details are known to me, that it constitutes the totality of processes by which I can come to accept arithmetical statements as true.*

Proof. For any given sound formal proof system S whose details I know, I can construct its Gödel formula $G(S)$, whose truth I thereby come to accept. Since $G(S)$ is not a theorem of S, the theorems of S do not exhaust the totality of arithmetical statements I can come to accept as true. □

So far, there is nothing to quarrel with here. But now Penrose further infers that the principles underlying the human mind are not computational in nature; and since they are presumed to be physical, the laws of physics must include some as yet unknown non-computational component.

What does Penrose mean by 'use'? Sometimes he refers to explicit procedures used by mathematicians, sometimes to low-level mechanisms in the brain. It seems obvious that the explicit procedures cannot exhaust the processes by which I come to believe arithmetical statements; while there is no reason to believe that the lower-level processes (which must include influences on the brain from outside) are, in detail, knowable to us.

Suppose that, against Penrose, we believe that all physical processes are computational in the sense that they can be simulated to an arbitrary degree of precision by means of digital computation (and hence expressible by means of a formal system of the kind to which Gödel's theorem applies). If in addition we believe with Penrose that all mental processes are, at some level of description, physical processes, then what we can actually deduce from Gödel's theorem is that *the totality of processes by which I can come to accept arithmetical statements as true is either unknowable to me, or unsound.* Penrose's argument depends on his finding this statement incredible; whereas to me it seems quite plausible. Deadlock!

Of course, Penrose's real target, which is Strong AI, doesn't come out of this too well either. We can hardly simulate what is unknowable to us!

7 Postscript on the Arithmetic of the Real Numbers

The set of real numbers $\mathbb{R}$ consists of all numbers expressible by means of finite or infinite decimal expansions. It includes the integers $\mathbb{Z}$ and the rational numbers $\mathbb{Q}$ and a whole lot else besides.

Many problems are *easier* to solve for this larger set than for the integers; indeed they can become trivial. Consider for example the following problem (a special case of Fermat's Last Theorem):

> Do there exist numbers x, y, z such that $x^7 + y^7 = z^7$?

If x, y, and z are required to be positive integers, this is quite a hard problem, and the answer turns out to be negative: there do not exist integers with this property. If, on the other hand, x, y, and z are allowed to be any real numbers, the problem is quite trivial. I can choose, say $x = y = 2$, and then put $z =^7 \sqrt{256}$, and I have a solution. Existence problems are easier to solve for real numbers than for integers because there are many *more* real numbers than integers. It is easier to see whether or not there is a real number with a specified property.

The first-order theory of the real numbers has the following non-logical vocabulary (with the usual interpretations):

> Constants: 0, 1
> Unary functions: $-$, $^{-1}$
> Binary function: $+$, $*$
> Binary predicate: $\leq$

The Polish logician Alfred Tarski (1902–1983) proved that this is a decidable theory. The following axioms are complete for this theory:

1. $x + (y + z) = (x + y) + z$
2. $x + y = y + x$
3. $x + 0 = x$
4. $x + (-x) = 0$
5. $x * (y * z) = (x * y) * z$
6. $x * y = y * x$
7. $x * 1 = x$
8. $x \neq 0 \rightarrow x * x^{-1} = 1$
9. $x * (y + z) = (x * y) + (x * z)$
10. $0 \neq 1$
11. $0^{-1} = 0$
12. $0 \leq x \vee 0 \leq (-x)$
13. $0 \leq x \wedge 0 \leq (-x) \rightarrow x = 0$
14. $0 \leq x \wedge 0 \leq y \rightarrow 0 \leq x + y$
15. $0 \leq x \wedge 0 \leq y \rightarrow 0 \leq x * y$
16. $x \leq y \leftrightarrow 0 \leq y + (-x)$
17. $\exists x \Phi(x) \wedge \exists y \forall x (\Phi(x) \rightarrow x \leq y) \rightarrow \exists z \forall y (\forall x (\Phi(x) \rightarrow x \leq y) \leftrightarrow z \leq y)$

Note that all function symbols have to be defined as total, and therefore 0^{-1} has to be assigned a meaning even though in reality the number 0 has no reciprocal; axioms 8 and 11 in effect define x^{-1} as the reciprocal of x when $x \neq 0$, and 0 otherwise. This does not interfere with the correct functioning of the system.

Axiom 17 (which is actually an axiom schema) expresses the least upper bound property of the real numbers: if a set of real numbers has an upper bound, then it has a least upper bound. As stated, the axiom does not cover all possible sets, since Φ can only be instantiated to first-order predicates constructible in the language; but since we are only considering the first-order theory of the reals, it does not matter that sets only expressible in higher-order logic are not handled.