Etica & Politica / Ethics & Politics, 2003, 1
http://www.units.it/~dipfilo/etica_e_politica/2003_1/15_monographica.htm

A Simple Exposition of Gödel's Theorem (*)

J.R. Lucas

Fellow of Merton College, Oxford
Fellow of the British Academy

In October 1997 I was asked to join in a discussion of the Gödelian argument at an undergraduate philosophy club in King's College, London; and I was asked to preface it with a very simple exposition of Gödel's (first) Theorem at a level at which first-year students could understand. Although there are many excellent accounts of Gödel's Theorem, I think it worth making accessible this very simplified account. It is over-simplified, but it may help readers to grasp the outline of the argument, before going on to fuller accounts.

Let me start with some autobiography. I was 17, a member of the school essay society, listening to a paper by the form atheist, who was putting forward a fairly abrasive version of reductive materialism. We were just blobs of protoplasm, bits of the primaeval slime that had evolved a bit, but still essentially a cocktail of carbon, hydrogen, oxygen and nitrogen. He was very persuasive. But then, I wondered, what did he think he was doing? Was he just trying to manipulate us, to program our nervous systems to go along with his opinions? No; he claimed to be in the right, to have reason on his side, to be pointing pout to us where the truth lay. But how could this be so, I asked, on his assumptions? If he was right, what we did, including what we said and what we thought, was completely determined by the material in our bodies and environment, and the laws of chemistry and physics that lay down what their movements and reactions must be.

It is an argument that has often come to me since, when I have heard people arguing against reason, and making out that all our apparent reasoning is merely the effect of our hormones, our infantile neuroses, or the economic interests of our class. In each case I counter that the very attempt to argue rationally belied materialism, Freudianism, or Marxism. It is a special case of a more general argument that has wide application in philosophy. When I started philosophy, Logical Positivism was very much in vogue, and my tutor tried hard to get me to believe the Verification Principle. So I asked him whether it was a tautology, perhaps showing some new meaning being given to `proposition', or `meaningful', so that only some selected instances could be accorded the favour of being thus described. He said No. Was the Verification Principle, then, an empirical proposition, discovered by careful research in which lots and lots of propositions had been examined, and none found to be meaningful except those that were analytic or empirical. He admitted, albeit a trifle reluctantly, that no such research had been carried out. In that case, I concluded triumphantly, the Verification Principle, if it were true, was itself meaningless, hoist by its own petard. He did not think it a very good argument, and told me to try harder to believe. But I thought it was a very good argument, and still think so, and regard it as one of the few arguments available to a philosopher when thinking about metaphysics, at a level of abstraction where most arguments fail to engage. See what it says about itself. Often it seems to be undercutting itself, cutting off the branch on which it is sitting, so to speak. It is silly to climb a ladder which is due to be thrown away when you have climbed it. Safer, and much more sensible to stay on the ground until one has found a destination which does not make out that really it is inaccessible.

I first heard of Gödel's Theorem in June 1948, when I had taken Maths Mods, and was wanting to change to Greats. I was told that there was some weird bit of mathematics involving prime numbers which was very strange. I suppose I had been trying to formulate my argument against materialism. Nine years later I was able to go to Princeton to study mathematical logic properly, and on my return tried out my argument on colleagues at Cambridge, then in a paper in 1959 to the Oxford Philosophical Society, which was finally published in 1961.

G□del's argument is self-referential. It is a development of the Epimenides paradox, known to St Paul. Epimenides was a Cretan and said that All Cretans were Liars. A modern version would be a backboard on which was written:

This statement is false

In this crude form we should not take it very seriously, any more than the ancients did Epimenides. One obvious flaw, pointed out by Gilbert Ryle, is that the `this' fails to refer properly. `This statement ': which statement? It refers to a statement that is still in course of being made. We don't know what the statement is that is being referred to until the statement has been completed, but it cannot be completed until the reference is made out properly. If we try to make sense of it, we get into a loop: `This statement, namely `This statement, namely `This statement, namely `This statement, namely . . . . .''''' Gödel circumvented Ryle's objection. He found a way of referring to well-formed formulae of formal logic which was independent of token-reflexive (or indexical) terms, such as `this'. Formal logic has relatively few symbols, and to each of these G□del assigned a number. He then could code a string of symbols by taking the odd prime numbers in order, and raising each to the power of the corresponding symbol. Thus if we want to refer to

(p V p) ---> p

and numbers assigned are

4 7 6 7 5 8 7

the number for the string is

$3^4.5^7.7^6.11^7.13^5.17^8.19^7$

This is an enormously big number, but it is just a number, and in principle we could refer to a well-formed formula by a single number. Instead of working out

$3^4.5^7.7^6.11^7.13^5.17^8.19^7$

let me pretend that it comes to 1729. Then it might be possible to write down, in a Ryle-proof manner

1729 - - - - - - - - - Well-formed formula no. 1729 is false

or, equivalently

1729 - - - - - - - - - Well-formed formula no. 1729 is not true

In order to do that, of course, we should need to be able to have all the terms in wff no. 1729 expressed in formal logic. Obviously `not' can be. Hence, Tarski argued, `true' cannot be, or we should be landed in a contradiction.

Formal logic cannot, on pain of inconsistency, contain a (meta-logical) predicate of well-formed formulae with the properties of `true'

or

`True' cannot be fully represented in formal logic

Tarski's Theorem

Formal logic is the language of computers. So a very simple version of my thought would be:
Computers cannot have a term with the properties of `true'.
Men do have a term with the properties of `true'. Therefore men are not computers.
If Pontius Pilate asks one of you `What is truth?' you can answer `I can't tell you, if I am only a machine---indeed, if I were, I should not be able even to understand the question.' (Unfortunately, when a tutor asks you, he is not disposed to think that you could tell him what truth is; nor even to understand the question.)
Instead of simply going for this negative conclusion, Gödel massaged truth, to represent it in formal logic so far as possible. Truth itself cannot be represented, but *provability-according-to-the-rules-of-formal-logic* can. What is a proof in formal logic? It is a sequence of well-formed formulae, starting with axioms, ending with the well-formed formula to be proved, with each successive step being a well-formed formula that follows from its predecessor according to some explicit rule of inference. So what Gödel needed to do was to code not just well-formed formulae---strings of symbols---but strings of well-formed formulae---strings of strings of symbols. But this can be done in essentially the same way as before, using this time not just the odd prime numbers, but 2 as well. Then a sequence of well-formed formulae can be expressed by an even number, and a putative proof of well-formed formula no.1729 would look something like
$2^{105}.3^{231}. . . . . 1873^{1729}$
To give a proper definition of a formal proof, Gödel needed to specify the axioms, and to formulate precisely the requirement that each well-formed formula was either an axiom of followed from earlier members of the sequence in virtue of one of the rules of inference. And having done this in meta-logical terms, he needed to show that granted his coding system, these requirements could be represented as arithmetically definable properties of numbers. It was a mammoth task, and took pages and pages of detailed working. But he managed it, and was able to define a relation between numbers which obtained just in case the first (even) number was the G□del number of a proof-sequence which was in fact a valid proof of the well-formed formula whose Gödel number was the second number in the relation. That is to say, there is a very complicated relation between numbers, Pr(x,y), which can be defined in terms of addition and multiplication, and holds when y is the Gödel number of a particular well-formed formula, and x is the Gödel number of a sequence of well-formed formulae which constitutes a proof of y. And then, generalising, he could have a general provability predicate which used the existential quantifier, and said just that there was a number which was the Gödel number of a sequence that was a proof of the well-formed formula whose Gödel number was the second number. So the well-formed formula, (Ex)Pr(x,1729) holds when there is a proof of the well-formed formula whose Gödel number is 1729, and conversely --(Ex)Pr(x,1729), holds when there is no proof of the well-formed formula whose Gödel number is 1729.
These two manoeuvres enable Gödel to refer to well-formed formulae by numbers, and to represent provability as a property of numbers definable in terms of the simple arithmetical operations of addition and multiplication, though the definitions are themselves very complicated. It remains to achieve *self*-reference, to find some Gödel number, 1729 as we have supposed for the sake of brevity, where wff no. 1729 turns out to *be* the wff --(Ex)Pr(x,1729) How can Gödel achieve this? He does it by means of a ``diagonalization operation'', like those used by Cantor to prove the non-denumerability of the continuum.
Cantor argues by *Reductio Ad Absurdum*. Suppose we could arrange all the real numbers between 0 and 1 in a denumerable list. The list would then look like
$0.a_{11}a_{12}a_{13}a_{14}a_{15}a_{16}a_{17}a_{18}a_{19} . . ..$
$0.a_{31}a_{32}a_{33}a_{34}a_{35}a_{36}a_{37}a_{38}a_{39} . . ..$

$0.a_{41}a_{42}a_{43}a_{44}a_{45}a_{46}a_{47}a_{48}a_{49}$ . . ..
$0.a_{51}a_{52}a_{53}a_{54}a_{55}a_{56}a_{57}a_{58}a_{59}$ . . ..
$0.a_{61}a_{62}a_{63}a_{64}a_{65}a_{66}a_{67}a_{68}a_{69}$ . . ..
$0.a_{71}a_{72}a_{73}a_{74}a_{75}a_{76}a_{77}a_{78}a_{79}$ . . ..
$0.a_{81}a_{82}a_{83}a_{84}a_{85}a_{86}a_{87}a_{88}a_{89}$ . . ..

where each of the a_{mn} is a digit 0,1,2,3,4,5,6,7,8,9.
Cantor then shows that there is a real number between 0 and 1 that has been left out, contrary to hypothesis.
Let b_{mm} be 1 if a_{mm} is 0, and 0 otherwise. Consider the number

$0.b_{11}b_{22}b_{33}b_{44}b_{55}b_{66}b_{77}b_{88}$ . . .

It cannot be first on the list because b_{11} is different from a_{11}, nor second because b_{22} is different from a_{22}, nor third, nor fourth, nor anywhere on the list because it will differ on the diagonal from that one. So not all the real numbers between 0 and 1 were on the original list, contrary to hypothesis.
Gödel used a similar technique to devise a well-formed formula which used a negative existential quantifier to say that a certain (very large) number did not have a carefully constructed arithmetical property, where the very large number turned out to be the Gödel number of the well-formed formula itself, and the carefully constructed arithmetical property was the one possessed by the Gödel numbers of wffs that were unprovable in the given system. The essential move is to take a two-place predicate, F(m,n), and then consider the case where m=n, thus converting the two-place predicate into a somewhat peculiar one-place one.
Let me wave my hand over the enormous amount of careful working needed to achieve this in a water-tight fashion, and claim that we have achieved self-reference.
wff no.1729- - - - - - - - - `wff no.1729 is unprovable in the system'

But then it must be true. Otherwise, if it were false, then wff no.1729 is not unprovable in the system, that is wff no.1729 *is* provable in the system; so the system is proving a false well-formed formula, and is fundamentally unsound. So provided the system---i.e. ordinary elementary arithmetic---is OK, it contains some well-formed formula, wff no.1729 as we have pretended, which is true and unprovable in the system.
I then gave a brief account of the central Gödelian argument against mechanism as in Minds, Machines and Gödel and Dr Gillies replied, putting forward Benacerraf's objection (God, the Devil and Gödel, *The Monist*, 1967, pp. 9-32.) that the account of one's mind might be undiscoverable. See D.A.Gillies, *Artificial Intelligence and Scientific Method*, Oxford, 1996, pp.142ff. and The Gödelian Argument: Turn Over the Page, pp.5-6; and gave his own ``political' account, as in D.A.Gillies, *Artificial Intelligence and Scientific Method*, Oxford, 1996, pp.151ff.
There is a much fuller, but still comprehensible, account in Ernest Nagel and James R. Newmann, *Gödel's Proof*, New York, 1958, London, 1959 Long ago, I found P.J Fitzpatrick, ``To Gödel via Babel'', *Mind*, LXXV, 1966, pp.332-350 very helpful.
There are now useful expositions in Roger Penrose, *The Emperor's New Mind*, Oxford, 1989; *Shadows of the Mind*, Oxford, 1994 which put forward the Gödelian argument against mechanism fully and carefully.

Notes

(*) For the copyright of the papers of J.R. Lucas see http://users.ox.ac.uk/~jrlucas/ **back**