Incompleteness and the possibility of AI

PHIL 30304
Jeff Speaks

November 29, 2018

1 Soundness and completeness

Roughly, the axioms of a theory are that theory's basic assumptions, and the theorems are the formulae provable from the axioms using the rules provided by the theory.

Arithmetic is a theory. It's axioms can be written like this:

> (P1) 0 is a number.
>
> (P2) The successor of any number is a number.
>
> (P3) No two numbers have the same successor.
>
> (P4) 0 is not the successor of any number.
>
> (P5) Any property which belongs to 0, and also to the successor of every number which has the property, belongs to all numbers.

These are called the Peano axioms. Using these axioms, we can define addition and other arithmetical operations, and prove, for example, that 56+9=65.

Claims that one can prove from the axioms of a theory are called *theorems* of that theory.

Suppose we are talking about the theory A of arithmetic. Then we can express the idea that a certain sentence p is a theorem of A — i.e., provable from A's axioms — as follows:

$$\vdash_A p$$

Now we can introduce the notion of the *valid* sentences of a theory. For our purposes, think of a valid sentence as a sentence in the language of the theory that can't be false.

Now we can ask, for any theory, how the theorems of the theory relate to its valid sentences. In the case of arithmetic, we said that 56+9=65 is a theorem of the theory. It is also valid; it can't be false. But in general we can ask two important questions of any theory:

- Are all of its theorems valid? If they are, we say that the theory is *sound*.
- Are all of its valid sentence theorems? If they are, we say that the theory is *complete*.

2 Gödel's incompleteness theorems

Gödel's incompleteness theorems are, as the name would indicate, proofs that certain mathematical and logical theories — one of which is Peano arithmetic — are not complete. That is, Gödel showed that there were certain valid sentences of those theories which were not provable from their axioms.

2.1 Gödel numbering and 'provable'

First we note that we can use the natural numbers to come up with 'names' for every formula of arithmetic. We do this by assigning every formula what is now called a Gödel number. There are many ways to do this. One is by first assigning a natural number to every basic symbol of the language of arithmetic, such as, for example, 0, 1, +, *, =, (, ... Then imagine that we have some formula of arithmetic, e.g.

0+1=1

We assign this formula a Gödel number by multiplying the Gödel number for the first digit of the formula (0) by 2, the Gödel number for the second digit of the formula (+) by 3, the Gödel number of the third digit (1) by 5, and so on, multiplying the Gödel number for the n^{th} digit by the n^{th} prime number. The sum of these n products is the Gödel number for the formula as a whole.

The point of using only prime numbers is that, this way, no two formulae will ever have the same Gödel number. The key points here are that every formula has a Gödel number, and that no two formulae every have the same Gödel number.

Rather than actually writing out the Gödel numbers for formulae, if we're talking about some sentence p, we'll use the following symbol for its Gödel number: $\#p$.

The point of this is to give us a way, using the language of arithmetic, of talking about the sentences of that language. You can think of the Godel number of a formula as a name for that formula.

Now assume that we can also define a predicate 'provable' such that, for the theory A of arithmetic,

$$\vdash_A p \iff \vdash_A \text{provable}(\#p)$$

i.e., 'provable($\#p$)' is a theorem of A if and only if 'p' is.

2.2 The fixed point lemma

Gödel showed that we could, for the theory A of arithmetic, find a sentence q such that

$$\vdash_A (q \iff \neg\, \text{provable}(\#q))$$

In other words, he showed that there was a formula such that it was provable in A that the formula was true iff it was not provable.

Suppose first that q is true. Then it follows from the above that it is not provable.

Suppose instead that q is false. We are assuming that theory in question is sound, and hence that falsehoods are never provable in the theory. But then it follows from the fixed point lemma that it is true. So the supposition that q is false leads to a contradiction, and q must be true.

This is the moral of the first incompleteness theorem: there is some formula in the language of arithmetic which is true and cannot be proven. This is sometimes called a 'Godel sentence.' These sentences are said to be 'undecidable' by the theory, since neither they nor their negations are provable from the axioms. In follows from the fact that there are such sentences that, in the terms we introduced earlier, arithmetic is incomplete.

2.3 Gödel's second incompleteness theorem

The second incompleteness theorem establishes that we cannot prove, in the language of arithmetic, the consistency of the axioms of arithmetic. The route this takes is a proof of the conditional claim

$$A \text{ is consistent} \rightarrow q$$

Since this conditional is provable, if the consistency of A were provable, 'q' would be provable as well. But we just saw above (in the proof of the first incompleteness theorem) that it isn't. So, the consistency of A is not provable in A.

3 Lucas on the consequences of the incompleteness theorems

The second incompleteness theorem says that one cannot prove, in the language of arithmetic, that arithmetic is consistent. And, more generally, for any language L which can state the truths of arithmetic, one cannot prove in L that L is consistent.

But, Lucas argues, this fact can be used to show that the mind is not a computing machine. Here is the key passage:

> 'Now a model of the mind must include a mechanism which can enunciate truths of arithmetic, because this is something which minds can do: in fact, it is easy to produce mechanical models which will in many respects produce truths of arithmetic far better than human beings can. But in this one respect they cannot do so well: in that for every machine there is a truth which it cannot produce as being true, but which a mind can. This shows that a machine cannot be a complete and adequate model of the mind. It cannot do everything that a mind can do, since however much it can do, there is always something which it cannot do, and a mind can. ... The Godelian formula is the Achilles heel of the cybernetical machine. And therefore we cannot hope to ever produce a machine that will do all that a mind can do ... '

Here is one way to think about the argument. Suppose (for reductio) that the mind is a computing machine, which knows things only by deriving theorems from axioms. Call the language which the mind is using L+.

The human mind can express the truths of arithmetic; so it follows from Godel's first incompleteness theorem that there are truths of L+ that are not provable in L+, and hence no knowable by the mind.

Further, it follows from Godel's second incompleteness theorem that (given the above) the mind cannot know that it is consistent.

But, Lucas argues, we can know that our own representation of the world is consistent. So, our minds cannot be machines, because we can know things that machines (provably) could not know.

What's the best reply? Does the fact that we can know whether arithmetic is consistent mean that we can always know when some formalized theory is consistent?