How Subtle is Gödel's Theorem
More on Roger Penrose

Martin Davis

According to Penrose 1990, none of his critics "have grasped the full import of the Gödelian argument ... " Applying "the Gödelization procedure" to an attempted automatization of the procedure as a whole would enable us to "obtain a new Gödel proposition that lies outside the scope of that automation. ... The Gödelian insight is a very slippery character ... "

I don't know how subtle and slippery all this is, but the technical side of what Penrose asserts is certainly correct. Indeed, it is a commonplace among professional logicians, and it is a little surprising that Penrose apparently believed that it could be news to working logicians like George Boolos and me. Iteration of the "Gödelization procedure" was first studied systematically in Alan Turing's doctoral dissertation (Turing 1939) and has been the subject of interesting research by Sol Feferman (1988) and others.[1] But it adds not one iota to Penrose's erroneous claim that Gödel's theorem can be used to show that mathematical insight can not be algorithmic.

There is really a problem in discussing these matters without being mercilessly technical. Whatever the "insights" involved, Gödel's remarkable *theorem* is a piece of technical mathematics. It is certainly worth discussing its philosophical implications, but without being carefully grounded in what the *theorem* actually states, it is all too easy to go astray. Therefore, I'm particularly pleased that Penrose proposes to discuss Gödel's theorem from an algorithmic point of view, since I believe that this can help considerably towards necessary demystification. He quite appropriately emphasizes that the Gödel sentences have the form $(\forall n)Q(n)$ where $Q(n)$ is a computable property of the natural number n and refers to such sentences as "statements of the form $[Q]$".[2] Again Penrose emphasizes that "we can directly apply the Gödelian insight to any algorithm that purports to be generating (say $[Q]$-type) mathematical truths." Thus, for any such algorithm $\mathbf{F}$, Gödel's *theorem* provides us with a corresponding $[Q]$-type sentence $\mathrm{G}(\mathbf{F})$ such that *if* $\mathbf{F}$ *is sound (in the sense that all the* $[Q]$*-type sentences it generates are true) then* $\mathrm{G}(\mathbf{F})$ *is true although it is not one of the sentences generated by* $\mathbf{F}$. Penrose goes on to say:

> The original algorithm is thus limited in what it is able to achieve: It is unable to incorporate the particular **insight** [emphasis mine - MD]

$$\text{“}\mathbf{F}\text{ sound”} \Rightarrow \text{“}\mathrm{G}(\mathbf{F})\text{” true}$$

[1] I was fascinated by this iteration when I began my graduate studies in 1948 and even thought of it as a possible dissertation topic.

[2] For reasons that need not concern us, logicians call assertions of this kind Π_1 sentences.

Calling this *fact* an *insight* rather than a *theorem* helps lead Penrose down the "slippery" path to mystification. In fact, "*theorem*" (an easy one at that) is what it is. For precision, the language of *computability theory*[3] is needed.[4] The only reason that **F** is "unable to incorporate" this "insight" is that as defined, **F** is limited to generating $[Q]$-type sentences, and this "insight" is (demonstrably) not a sentence of that class.[5] However, it is a quite ordinary sentence of elementary number theory and can be proved with no difficulty whatever in any formal system adequate for elementary number theory, such as for example Peano arithmetic. Note that this powerful form of Gödel's theorem applies uniformly to any formalism whatever. Thus, in particular, *it is demonstrable in Peano arithmetic that the soundness of the Zermelo-Fraenkel axioms (or of any of their popular extensions) implies the Gödel sentence corresponding to that system of axioms.*

Penrose objects to the claim by his critics that "the Gödelization procedure could itself be automated," arguing that "this is not so if by 'procedure' we mean the above 'insight'." Since this "insight" is but a humble theorem, formalizable in all but the weakest logical systems, it is hard to see what Penrose could have had in mind.

Penrose claims that his argument was "certainly appreciated by Gödel himself in the 1930s and never properly refuted since ..." In evidence, he cites the published two volumes of Gödel's collected works in their entirety (Gödel 1986/1990). Now despite disagreements, I enjoyed *Emperor* thoroughly; however I was irritated by Penrose's slapdash scholarship, for example by his conflation of Hilbert's tenth problem with Hilbert's Entscheidungsproblem.[6] But this is really too much. If Gödel had "certainly appreciated" these matters, why not cite his statements? The one specific citation is from Rudy Rucker's popularization (1982, p. 171). Rucker quotes Gödel's remarkable (but still at this writing, unpublished[7]) Christmas 1951 Gibbs lecture invited by the American Mathematical Society. And, what is really curious is that part of what Rucker quotes Gödel as saying directly contradicts Penrose's position:

[3] Also called *recursion theory*.

[4] Here's the proof: $[Q]$-type sentences can be coded in a natural way by integers, e.g. using the codes of Turing machines carrying out the algorithm to compute Q. Let T be the set of codes of those sentences of the class that are *true*. The algorithm **F** can be thought of as generating a recursively enumerable set T_0 of such codes. To say that **F** is *sound* is just to say that $T_0 \subseteq T$. The set T is not recursively enumerable, and therefore it can't be identical to the recursively enumerable set T_0. So, there is a number q_0 in the set $T - T_0$. This number q_0 must code a true $[Q]$-type sentence not obtainable from the algorithm **F**. (For details, see for example Davis 1958/1982, 117-122.)

An additional technical point is that the set T is what is called *productive*. This means that there is an actual algorithm which, given a particular Turing machine defining a formal system **F** as input, will output the corresponding number q_0, that is it will output a true $[Q]$-type sentence not provable by **F**.

[5] A simple calculation shows that it is what logicians call a Δ_2 arithmetic sentence.

[6] It was not only that Hilbert's posing of these problems was separated by a third of a century; I personally spent twenty years of my life working on Hilbert's tenth problem at a time when the Entscheidunsproblem had long since been settled!

[7] It will appear soon in the third volume of Gödel's collected works. I was lucky enough to have heard Gödel give the lecture and was strongly influenced by it.

> On the other hand, on the basis of what has been proved so far, it remains possible that there may exist (and even be empirically discoverable) a theorem-proving machine which in fact is equivalent to mathematical intuition, but cannot be *proved* to be so, nor even be proved to yield only *correct* theorems of finitary number theory.

In fact, during the 1930s Gödel was ultra careful to avoid controversial statements and to stick to what could be proved. However, the Gibbs lecture was really a bombshell. Gödel argued forcefully that his theorem had important philosophical implications. But, as the above quotation makes clear, he never asserted that mathematical intuition could be shown to be non-algorithmic. Instead, his argument, in the form of a dilemma, was truly subtle. Either the human mind is mechanical, and hence equivalent to a Turing machine, and then there will be an absolutely undecidable statement of type $[Q]$, or else there is a non-mechanistic aspect to the mind, and we must accept a vitalist position. In either case, said Gödel, we are led to a Platonist position regarding the subject matter of mathematics. I found and continue to find much in this argument that is persuasive, and was astonished to note Penrose's statement about me that I am a "mathematician of a formalist disposition." Gödel understood very well, what Penrose seems to have missed, that a Platonist position concerning mathematical entities is perfectly consistent with a mechanist view of mind.

Let me be very clear about this: I think it quite likely that Gödel would have agreed with Penrose's judgement that mathematical intuition can not be the product of an algorithm. In fact Gödel apparently believed that the human mind could not even have been the product of natural evolution. However, Gödel did not assert that such conclusions were a consquence of his famous theorem.

It all comes down to the issue of the *soundness* of the formal systems that are constructed to encapsulate mathematical thinking. However it is that we come to accept the soundness of such a system, Gödel's beautiful theorem compels us to also accept the soundness of a stronger system as well. Gödel himself believed that we truly *see* concepts, as with a sense organ, and that as our perception clarifies, we can formulate axioms expressing the properties of these concepts, and which we can therefore *see* to be true. But he never claimed that this point of view was a necessary consequence of his incompleteness theorem. Personally, I find Gödel's objective metaphysics hard to swallow. My own view is that confidence in a powerful formal system develops over the course of time, very much as in the case of a physical theory. But that is another question. To repeat what I said in my first commentary on Penrose (Davis 1990): the question of whether mathematical insight is the product of an algorithmic process is an important and difficult question. But Gödel's theorem has nothing to say about it.

One last point: Penrose expresses disappointment over the relative lack of discussion of his physical ideas. However, although he included my name in the list of those who do say something about the physics, very much to my disappointment, he said nothing about my specific comment. Although it is not central to his concerns, Penrose makes it clear [*Emperor*, p. 288] that he finds something not entirely satisfactory about the procedure in which the process of "quantizing" a classical theory involves replacing real observables by operators. I truly believe that, based on an important insight of the contemporary logician Takeuti, I have contributed to an understanding of this issue (Davis 1977). I would be most pleased if Roger Penrose were to comment on this.

BIBLIOGRAPHY

Davis M. (1958/1982) *Computability and unsolvability,* McGraw-Hill. REPRINTED WITH AN APPENDIX Dover.

(1977) A relativity principle in quantum mechanics. *International Journal of Theoretical Physics* 16:867-74.

(1990) Is mathematical insight algorithmic? *Behavioral and Brain Science.* 13:659-60.

Gödel, K. (1986/1990) *Kurt Gödel, collected works,* vol. I (orig. pub. 1929-1936) & vol. II (orig. pub. 1938-1974), ed. S. Feferman, J. W. Dawson Jr., S. C. Kleene, G. H. Moore, R. M. Solovay, J. van Heijenoort. Oxford Universit Press.

Feferman S. (1988) Turing in the land of O(Z). *The universal Turing machine - a half-century survey,* Rolf Herken, editor, 113-147, Verlag Kemmerer & Unverzagt, Hamburg, Berlin/Oxford University Press.

Penrose, R. (1990) Author's response: the nonalgorithmic mind. *Behavioral and Brain Science.* 13: 692 - 705.

Rucker, R. (1984) *Infinity and the mind: The science and philosophy of the infinite.* Paladin Books/Granada Publishing Ltd.

A. Turing (1939/1965) Systems of logic based on ordinals. *Proc. London Math. Soc., ser. 2* 45:161-228. REPRINTED: *The Undecidable,* M. Davis, ed., 155 - 222. Raven Press, New York.