Penrose's Gödelian argument

Solomon Feferman
Department of Mathematics
Stanford University
Stanford, CA 94305-2125

Penrose redux

In his book *Shadows of the Mind: A search for the missing science of consciousness* [*SM* below], Roger Penrose has turned in another bravura performance, the kind we have come to expect ever since *The Emperor's New Mind* [*ENM*] appeared. In the service of advancing his deep convictions and daring conjectures about the nature of human thought and consciousness, Penrose has once more drawn a wide swath through such topics as logic, computation, artificial intelligence, quantum physics and the neuro-physiology of the brain, and has produced along the way many gems of exposition of difficult mathematical and scientific ideas, without condescension, yet which should be broadly appealing.[1] While the aims and a number of the topics in *SM* are the same as in *ENM*, the focus now is much more on the two axes that Penrose grinds in earnest. Namely, in the first part of *SM* he argues anew and at great length against computational models of the mind and more specifically against any account of mathematical thought in computational terms. Then in the second part, he argues that there must be a scientific account of consciousness but that will require a (still to be found) non-computational extension or modification of present-day quantum physics.

[1] Take, as just one example, the vivid mini-"history" in *SM*, pp. 249–256, of the origins of probability theory and complex numbers in the work of the 16th century mathematician and physician, Gerolamo Cardano — as a prelude to an explanation of Schrödingerian quantum mechanics.

I am only competent to say something substantive about the first part of the new effort, resting as it does to a considerable extent on a version of Gödel's (first) incompleteness theorem. Penrose had advanced that previously in *ENM*, but the line of argument was much criticized, as it had been in the past when advanced by others (e.g. J.R. Newman and E. Nagel, and J.R. Lucas)[2]. So now Penrose has gone to great lengths in *SM* to lay out his Gödelian argument and to try to defend it against all possible objections. I must say that even though I think Gödel's incompleteness theorems are among the most important of modern mathematical logic and raise fundamental questions about the nature of mathematical thought, and even though I am personally convinced of the extreme implausibility of a computational model of the mind, Penrose's Gödelian argument does nothing for me personally to bolster that point of view, and I suspect the same will be true in general of similarly inclined readers. On the other hand, I'm sure that those whose sympathies lie in the opposite direction will find reasons to dismiss the Gödelian argument quickly on one ground or another without wading through its painful elaboration. If I'm right, this is largely a wasted effort – diligent as it is. Nevertheless, it's there, and I feel obliged to address at least parts of it, especially its more technical aspects.

While I have disavowed competence concerning Part II of *SM*, I can't help registering my impression that the effort there is entirely quixotic. What Penrose aims to do is substitute one "nothing but" theory for another: in place of "the conscious mind is nothing but the manifestation of sub-atomic physics". Can we really ever expect a completely reductive theory of one sort or another of human cognition? Surely, no one theory will serve to "explain" the myriad aspects of this phenomenon. As with any other scientific study of human beings – inside and out – such an enterprise will continue to need to bring to bear psychology, psycho-physics, physiology (neuro- and otherwise), biochemistry, molecular biology, physics (macro- and micro-) and lots of stuff in between (including computational models of all kinds). In my opinion Penrose's "missing science of consciousness" is a mirage.

[2]For earlier critical discussion, cf. the collection Anderson (1964). For criticism in various of the peer commentaries on *ENM* (with responses by Penrose), cf. *Behavioral and Brain Sciences* v. 13 #4 (1990), 643–705, and v. 16 # 3 (1993), 611–622.

The logical facts

While Penrose's formulation of Gödel's theorem is by itself unexceptionable, his subsequent discussion of it – especially in relation to Gödel's won formulation and various of it generalizations – is unfortunately marred by a number of errors. I assume here some familiarity with mathematical logic and the relevant material from Kleene (1952); the reader who does not have that familiarity should skim the following before proceeding to the next section of this review. Unless otherwise indicated, pagination or section references (e.g. '2.5') are to *SM*.

Penrose's form of Gödel's incompleteness theorem is stated in terms of Turing machine computations as follows (pp. 74–75):

Theorem 1. Suppose A is a Turing machine which is such that whenever A halts on an input (q, n) then $C_q(n)$ does not halt. Then for some k, $C_k(k)$ does not halt, yet A does not halt on (k, k). In other words, if the halting of A is a sufficient condition for the non-halting of Turing machines then it is not a necessary condition for that; still more briefly: soundness of A implies incompleteness of A.

The proof of Theorem 1 is just a variant of the standard diagonal argument, originating with Turing in 1937, that the halting problem for Turing machines is not effectively decidable. As a form, though, of Gödel's incompleteness theorem, it is very close to Kleene's generalized form of that result, established in 1943 and exposited in Kleene (1952) p. 302 as Theorem XIII. That make use of a very general notion of formal system F, the main condition for which is that the set of "provable formulas" is effectively enumerable. Suppose in particular that F contains effectively given "formulas" $\phi(\mathbf{q}, \mathbf{n})$ which are supposed to "express" the predicate $P(q, n)$ which holds just in case $C_q(n)$ does not halt. F is said to be sound or correct for P if whenever F proves $\phi(\mathbf{q}, \mathbf{n})$ then $P(q, n)$ holds, and it is said to be complete for P if the converse is true. In slightly weakened form, Kleene's theorem (loc. cit.) is then as follows:

Theorem 2. If F is a formal system (in the general sense) which is sound for the predicate P then it is not complete for it. In particular, there is a k such that $C_k(k)$ does not halt though F does not prove $\phi(\mathbf{k}, \mathbf{k})$.

Assuming Church's Thesis, Theorem 2 follows Theorem 1, since every re-

cursively enumerable set of pairs (q, n) is the same as the set of inputs on which Turing machine halts. Conversely, to obtain Theorem 1 from Theorem 2, simple take the "formula" $\phi(\mathbf{q}, \mathbf{n})$ to be the pair (q, n) and the set of "provable formulas" of F to be the set of pairs on which A halts.

We must now examine the relationship of these results with the usual formulation of Gödel's incompleteness theorems. Here we deal with formal systems in the logical sense, i.e. systems F whose formulas are built up from basic arithmetical (and possibly other) relations by means of the propositional connectives (such as $\neg, \wedge, \vee, \rightarrow$) and quantifiers (such as $\forall, \exists$) and whose provable formulas are obtained from a given set of axioms (both logical and non-logical) by closing under certain rules of inference. Moreover, F is assumed to be effectively given, i.e. the set of axioms of F and its rules of inference are supposed to be effectively decidable, so that its set of provable formulas is effectively enumerable. Finally, F is supposed to be "sufficiently strong", i.e. contain a modicum F_0 of elementary number theory (or arithmetic). Over the years, the statement of Gödel's incompleteness theorems has been steadily strengthened by a steady weakening of what is assumed for F_0. In 1931, Gödel had taken it to be a version of simple type theory over a number-theoretical base, but he soon weakened that to a form of the first-order system of Peano Arithmetic PA. Subsequently, for Gödel's first incompleteness theorem, this was further weakened considerably by R.M. Robinson's fragment Q of arithmetic, and for the second incompleteness theorem to the subsystem Σ_1-IA of PA based on induction applied only to Σ_1 formulas. For details, cf. Gödel's 1931 and related papers in the original and in translation, with an introductory note by Kleene, in Gödel (1986), pp. 126 ff, and the expositions in Kleene (1952), pp. 204–213 and Smorynski (1977). For simplicity, we assume throughout the following that F_0 is PA, and that F is an (effectively given) formal system in the logical sense which contains PA.

As Penrose notes, the class of Π_1 formulas is of special significance in connection with the Gödel theorems. These are ones of the form $\forall x R(x)$, where R expresses an effectively decidable (= recursive) property of the natural numbers and the intended range of 'x' is the set of natural numbers. Dual to this class is the class Σ_1 of formulas of the form $\exists x S(x)$ where S is effectively decidable; in classical logic, these are equivalent to negations $\neg\forall x\neg S(x)$ of Π_1 formulas. We may consider similarly Π_1 and Σ_1 formulas with free variables such as $\forall y R(x, y)$ and $\exists y S(x, y)$ with decidable R, S, resp.

The following examples are particularly relevant to the Gödel theorems: (i) We have a decidable relation $\text{Proof}_{\text{F}}(x, y)$ which express that y is (the Gödel number of) a proof in F of the formula (with number) x; then (ii) the Σ_1 formula $\text{Prov}_{\text{F}}(x) := \exists y\ \text{Proof}_{\text{F}}(x, y)$ expresses that the formula (with number) x is provable in F, while the Π_1 formula $\forall y \neg \text{Proof}_{\text{F}}(x, y)$ expresses that x is not provable in F; in particular, (iii) if c is the number of the formula $\mathbf{0} = \mathbf{1}$, the Π_1 formula $Con(\text{F}) := \forall y \neg \text{Proof}_{\text{F}}(\mathbf{c}, y)$ expresses that F is consistent. Next, (following Kleene), we have (iv) a decidable relation $T(z, x, y)$ which expresses that y is (the number of) a terminating computation at input x on the Turing machine C_Z; so (v) the Σ_1 formula $\exists y T(z, x, y)$ expresses that $C_Z(x)$ halts, and (vi) the Π_1 formula $\forall y \neg T(z, x, y)$ expresses that $C_Z(x)$ does not halt; in particular, (viii) for each k, the Π_1 sentence $\forall y \neg(\mathbf{k}, \mathbf{k}, y)$ expresses that $C_k(k)$ does not halt.

The system F is said to be *sound* for a class $\mathcal{S}$ of sentences if whenever F proves ϕ with ϕ in $\mathcal{S}$ then ϕ is *true* in the structure $\mathbb{N}$ of natural numbers; F is said to be *complete* for the sentences in $\mathcal{S}$ if the converse holds. F is said to be ω-*consistent* if there is no formula $\phi(x)$ such that F proves $\neg\phi(\mathbf{n})$ for each natural number n, and yet F proves $\exists x \phi(x)$. It is said to be 1-*consistent* if this condition holds for decidable ϕ. It is obvious that if F is ω-consistent then it is 1-consistent, and that in turn implies that it is consistent, since an inconsistent system proves all formulas. Under our general assumption that F is sufficiently strong (i.e. contains PA as a subsystem), we have:

Lemma 3.
(i) F is complete for Σ_1 sentences.
(ii) F is 1-consistent if and only if F is sound for Σ_1 sentences.
(iii) F is consistent if and only if F is sound for Π_1 sentences.

The idea for the proof of (ii) is that if F is 1-consistent and proves $\exists x S(x)$ where S is decidable then there must exist an n such that $S(n)$ holds, since otherwise we could prove $\neg S(\mathbf{n})$ for each n. The converse is immediate by definition. The idea for the proof of (iii) is that if F is consistent and proves $\forall x R(x)$ with R decidable, then for each n, F proves $R(\mathbf{n})$, hence $R(\mathbf{n})$ must be true, for otherwise $\neg R(\mathbf{n})$ would be provable in F as a special case of (i).

The result (iii) of Lemma 3 is a fundamental observation due to Hilbert, since it shows that any success of his consistency program for a system F would establish the correctness of F for the "real" (i.e. Π_1) sentences.

For his completeness results, Gödel constructed a Π_1 sentence $G(\mathrm{F})$ equivalent (in PA) to $\forall y \neg \mathrm{Proof}_{\mathrm{F}}(\mathbf{g}, y)$, where g is the Gödel number of $G(\mathrm{F})$. Thus $G(\mathrm{F})$ provably expresses of itself (via its Gödel number) that is is not provable in F.

Theorem 4. (Gödel's 1st incompleteness theorem).
(i) If F is consistent then $G(\mathrm{F})$ is not provable in F.
(ii) If F is ω-consistent then $\neg G(\mathrm{F})$ is not provable in F.

Now the hypothesis in (i) is equivalent by Lemma 3 to the soundness of F for Π_1 sentences, and since under this hypothesis the sentence $\forall y \neg \mathrm{Proof}_{\mathrm{F}}(\mathbf{g}, y)$ is true and hence $G(\mathrm{F})$ is true, we conclude from the *first part* of the 1st incompleteness theorem that:

Corollary 5. If F is sound for Π_1 sentences then F is not complete for them.

In this form, the first part of Gödel's 1st incompleteness theorem is of the same character as Theorem 1 (Turing-Penrose) as well as Theorem 2 (Kleene). Note that the hypothesis of the second part of Theorem 4 can be replace immediately by the assumption that F is 1-consistent. For if $\neg G(\mathrm{F})$ is provable in F then so also is $\exists y \mathrm{Proof}_{\mathrm{F}}(\mathbf{g}, y)$. But that sentence is false by the first part of the theorem. Note also that if $\neg G(\mathrm{F})$ is added to F as an axiom, and F is consistent, the resulting system is still consistent (by the first part of Theorem 4) but not 1-consistent, hence not ω-consistent.

In 1937, Rosser constructed a Π_1 sentence $R(\mathrm{F})$ which is such that if F is consistent neither $R(\mathrm{F})$ nor $\neg R(\mathrm{F})$ is provable in F. However, $R(\mathrm{F})$ is less useful than $G(\mathrm{F})$, as the following shows:

Theorem 6. (Gödel's 2nd incompleteness theorem).
(i) PA proves $Con(\mathrm{F}) \leftrightarrow G(\mathrm{F})$.
(ii) Hence, if F is consistent then F does not prove $Con(\mathrm{F})$, i.e. F does not prove its own consistency.

The idea of a proof of $Con(\mathrm{F}) \rightarrow G(\mathrm{F})$ in (i) is to formalize in PA the proof of the first part of the 1st incompleteness theorem. The converse, that $G(\mathrm{F}) \rightarrow Con(\mathrm{F})$, is trivial, since if any sentence is not provable in F then $\mathbf{0} = \mathbf{1}$ is not provable in F.

After this extensive, but, as we shall see, necessary excursus, we can finally return to Penrose's *SM*.

pp. 74–75 (sec. 2.5). The idea of a computational procedure A being *sound* is explained here by the statement that if A halts on input (q, n) then $C_q(n)$ does not halt. As we have seen, the conclusion is equivalent to the Π_1 sentence $\forall y \neg T(\mathbf{q},\mathbf{n}, y)$, and soundness of A is a special case of soundness of a formal system for Π_1 sentences. However:

pp. 90–92 (sec. 2.8). In Penrose's account of Gödel's incompleteness theorem, he says (p. 91) that if a formal system is *sound* then "it is certainly ω -consistent". This is a *different* notion of soundness from that on pp. 74–75, since ω-consistency is stronger than consistency, i.e. than soundness for Π_1 sentences. Penrose does not explain here what is meant by this new notion of soundness, but implicit in what he says is soundness for *all* (arithmetical) sentences [cf. the discussion of p. 112 below]. At any rate, the notion of soundness required for Penrose's further discussion is ambiguous between that of pp. 74–75 and that of p. 91.

Next on p. 91, Penrose introduces the notation 'Ω(F)' for the [formal] assertion that the system (F) is ω-consistent. He says that "Gödel's famous incompleteness theorem tells us that Ω(F) is *not a theorem* of F... provided that F is actually ω-consistent." As we have seen, Gödel's 2nd incompleteness theorem tells us that Con(F) is not a theorem of F provided F is simply consistent, *a fortiori* Ω(F) is not a theorem of F under the same conditions. The hypothesis of ω-consistency of F (or its weakening, 1-consistency) is needed only if we want also to conclude that $\neg Con$(F) (or equivalently, $\neg G$(F)) is not a theorem of F.

Penrose further says here that he will use the notation 'G(F)' for the [formal] assertion that F is consistent. He then says that Rosser's theorem tells us that if F is consistent then G(F) is not a theorem of F; but that is what Gödel's 2nd incompleteness theorem tells us, not Rosser's. Penrose further muddies the picture by saying that he will "not bother to draw a clear line between consistency and ω-consistency" in most of his discussions, but that "the version of the Gödel theorem that I [Penrose] have actually presented in sec. 2.5 is essentially the one that asserts that if F is ω-consistent, then it cannot be complete, being unable to assert Ω(F) as a theorem." Instead,

as we have seen, what he showed in sec. 2.5 is a version of the first part of Gödel's incompleteness theorem, that if F is sound for Π_1 sentences [$\Leftrightarrow$ consistent] then G(F) is not a theorem of F.

On p. 92, Penrose says that in order to discuss the actions of Turing machines, F must contain the minimum operator (μ) symbol. It is true that, in Kleene's normal form theorem, the value of $C_q(n)$ is of the form $U\left(\mu y.T\left(q,n,y\right)\right)$ when $\exists yT(q,n,y)$ holds; but the statement of halting or non-halting of $C_q(n)$ does not require the explicit presence of μ among the symbols of F.

p. 96. It is stated here that "...both Ω(F) and G(F) are Π_1 sentences." This is correct for G(F) but not for Ω(F) which is, instead, a sentence in Π_3 from, i.e. of the form $\forall x\exists y\forall zR(x,y,z)$ with R decidable [work it out]. Even 1-consistency is a Π_2 sentence, which is not equivalent to a Π_1 sentence.

p. 108. Penrose says here that if F* and F** are obtained from F by adjoining G(F) and $\neg G$(F) resp. as axioms, and if F is consistent then F* and F** are both consistent. This is correct for F** by ordinary logic, but not for F*. The following is a counter-example: let F be obtained from PA by adjoining $\neg G$(PA) or, equivalently, $\neg Con$(PA) as an axiom. Thus F is PA** in Penrose's notation, and so F is consistent. But in this case, since F* includes Con(F) and PA is contained in F, we have that F* proves Con(PA), so F* is inconsistent. What is needed to insure that F* is consistent is the assumption that F is 1-consistent (which is *not* the case for F=PA**); as it happens, it can be shown that if F is 1-consistent so also is F*.

pp. 109–110. The discussion of the non-categoricity of the first-order version of PA of Peano's axioms vs. the categoricity of the second-order version of those axioms is misleading since it lumps together first-order quantification with second-order quantification. What the latter does is allow one to quantify over properties P in the induction axiom, namely as $\forall P[P(0) \wedge \forall x(P(x) \rightarrow P(Sx)) \rightarrow \forall xP(x)]$. However, Penrose is right in saying that for this second-order axiom to guarantee categoricity we need to regard it semantically, i.e. to interpret the variable 'P' in '$\forall P$' as ranging over arbitrary subsets of the first-order domain on interpretation, and there is no effective formal system complete for this semantics (by Gödel's incompleteness theorem).

p. 112. In the discussion of **Q18** it is asserted that we cannot "properly encapsulate 'soundness' or 'truth' within any formal system – as follows by a famous theorem of Tarski". This settles definitely the earlier ambiguity between the notions of soundness used on pp. 74–75 and that of p. 91, i.e. here soundness is taken as truth of all sentences (at least all arithmetical ones); then Tarski's theorem on the non-definability of truth certainly applies provided the system F under consideration is consistent. Penrose goes on to say that for restricted notions of soundness we can prove in F, or even PA, that if F is sound then G(F) holds. In particular, he says that PA proves Con(F)$\rightarrow$ G(F). This is strange, because on p. 91 he said that he will use 'G(F)' for the formal statement that F is consistent, i.e. for 'con(F)': but for that identification the implication is trivial. The implication Con(F)$\rightarrow$ G(F) is only of interest if one takes G(F) to be Gödel's sentence that expresses of itself that it is not provable in F (cf. Theorem 6(i) above). The next strange statement on p. 112 is that one can prove that 'F ω-consistent' implies 'Ω(F)', since on p. 91 Penrose defined Ω(F) to be the formal statement of the ω-consistency of F; on that identification the implication is once more trivial.

p. 114. The description of my results on Turing's ordinal logics is incorrect. First of all, the reference given is to Feferman (1988), which contains a historical exposition of Turing's seminal work (1939) and subsequent work on this subject (under the new name, transfinite recursive progressions of formal systems). The appropriate reference for my own original work there should have been Feferman (1962). It was Turing (not me) who showed in his 1939 paper that the ordinal logic obtained by iteration of adjunction of consistency statements starting with PA and proceeding through the recursive ordinals is complete for Π_1 statement (in fact at a surprisingly low level); Turing had hoped to improve this to completeness for Π_2 sentences. In my 1962 paper I proved that: (i) Turing's ordinal logic is incomplete for Π_2 sentences; (ii) the same holds for progressions based on transfinite iteration of the so-called local reflection principle; (iii) but one obtains completeness for *all* arithmetical sentences in a progression based on the transfinite iteration of the so-called global or uniform reflection principle. However, the following comments by Penrose about the significance of Turing's and my work are correct: "...there is no algorithmic procedure that one can lay down beforehand which allows one to do this systematization for *all* recursive ordinals once and for all", and that "...repeated Gödelization...does not provide us with a mechanical

procedure for establishing the truth of Π_1 sentences."

I have not detailed all the occurrences of technical errors that Penrose makes in connection with Gödel's incompleteness theorems in Ch. 2, many of which also propagate through Ch. 3. Given the weight that Penrose attaches to his Gödelian argument, all these errors should give one pause. One has here lots more of the "slapdash scholarship" that Martin Davis complained about in his commentary on *ENM* (1993) p. 116, and they suggest that he may stretch that scholarship perilously thin in areas distant from his own expertise. The main question, though, is whether these errors undermine the conclusions that Penrose wishes to draw from the Gödelian argument. I don't think that they do, at least *not by themselves*. That is, I think that the extended case he makes from sec. 2.6 on through the end of Ch. 3 would be unaffected if he put the logical facts right; but the merits of that case itself are another matter.

What follows from Gödel's incompleteness theorem?

Here I shall be less systematic in tracking Penrose. It must be emphasized again that what his case really rests on is the first half of Gödel's 1st incompleteness theorem (Theorem 4(i) above)– that if a suitably strong formal system F is consistent then the Π_1 sentence $G(\mathrm{F})$ is not provable in F – combined with Hilbert's observation (Lemma 3(iii) above) that F is consistent if and only if F is sound for Π_1 sentences. Finally, we have by Gödel's 2nd incompleteness theorem (Theorem 6 above) that $G(\mathrm{F})$ is equivalent in a base system (e.g. PA) to $Con(\mathrm{F})$. The ω consistency of F and statement $\Omega(\mathrm{F})$ are simply red herrings for Penrose's argument and should be ignored. The reformulation of incompleteness in terms of Turing machines in sec. 2.5 is of course important if one is to argue that mathematical thought is not mechanical, but it is *just* a reformulation as Penrose brings out: every theorem-generating machine can be recast as a formal system and vice-versa. However, it is the model of mathematical thought in term of formal systems that is closer to the nature of that thought itself, i.e. to its concepts and modes of reasoning. What is misleading in the equivalence between Turing machines and formal systems is the way theorems are actually obtained in the working experi-

ence of mathematicians On the algorithm model, one starts with an input (q, n) on machine A in an effort to establish that $C_q(n)$ does not halt, i.e. one starts with the "statement" possibly to be established and plugs away mechanically following the algorithm that determines A in the hopes that it will end by "proving it". The analogue for a formal system F would be to start with a statement ϕ, possibly to be established, and mechanically generates, one after another, all proofs in F, looking to see if one of them ends with ϕ. But it would be ridiculous to think that anything like such a search through proofs takes place in the activity of working mathematicians. How is it that they actually arrive at proof is through a marvelous combination of heuristic reasoning, insight and inspiration (building, of course, on prior knowledge and experience) for which there are no general rules, though some patterns have been discerned by Pólya and others: these is *no* formula for mathematical success. It is only when one finally arrives at a proof that one can check (mechanically, in principle, but not in practice) that it does indeed establish the theorem in question. So on the face of it, mathematical thought as it is actually produced is not mechanical; I agree with Penrose that in this respect, *understanding* is essential, and it is just this aspect of actual mathematical thought that machines cannot share with us. Beyond that, his entire drive is to nail down this conviction by showing that mathematical thought cannot even be *re-represented* in mechanical terms, as a result of the Gödel theorem. In my view, instead of increasing this conviction, this effort raises more questions than it answers and leads one off into dead-end dialectics. Here are some reasons.

Penrose begins by stating as the main conclusion $\mathcal{G}$ from the Gödel-Turing incompleteness theorem: "Human mathematicians are not using a knowably sound algorithm in order to ascertain mathematical truth" (p. 767). More specifically, in terms of formal systems: if mathematicians can come to know that a system F is sound, then F cannot be used to ascertain the truth of the true Π_1 statement $G(\mathrm{F})$. Now, as we have noted, there is an ambiguity in Penrose's use of the notion of soundness between that for Π_1 sentences and that for all sentences. All that the Gödel incompleteness theorem requires of F is the former, since that is equivalent to the consistency of F. But Penrose tends to emphasize the global notion of soundness and to tie it to his Platonistic philosophy of mathematics. The argument goes something as follows: how could we know that F is sound if we did not understand what F is about – its intended interpretation – and see that the axioms of F

are all *true* of that interpretation and that its rules of inference all *preserve truth*? It is by such means, the argument continues, that we recognize the soundness of systems from PA all the way up to ZF set theory and beyond. And once we recognize the soundness of a system F and accept it as part of the principles on which we can rely, we see that $G(\mathrm{F})$ is true and must accept it too, and so by Gödel's theorem, we are required to accept something that goes beyond F.

Two problems with this argument are that, first of all, there may be other ways of recognizing the truth of $G(\mathrm{F})$ than through a global notion of truth for F, and secondly, the assumption of an intended interpretation for set-theoretical formalisms is highly problematic. The first is what is achieved by proof theory. While it is generally agreed that Hilbert's program to establish the consistency of stronger and stronger formal systems by purely finitary proof-theoretical methods cannot be carried through as a result of Gödel's 2nd incompleteness theorem, a *relativized* form of Hilbert's program has been successful by these means (cf. Feferman (1988a)). Relativized proof theory yields verification of the consistency of a system F by reduction to the consistency of another system F', and progress is achieved thereby when one has more compelling reasons for accepting F' than F to begin with. In particular, various *prima-facie* non- constructive systems have been reduced in this way to constructive systems, and systems of analysis based on impredicative set-existence principles have been reduced to predicative systems. Indeed, it has been shown that the bulk of everyday mathematics can be formalized in such relatively weak systems, and it appears that all of scientifically applicable mathematics can be formalized in a system which is proof-theoretically reducible to PA (cf. Feferman (1993)). While mathematicians may *conceive* of what they are talking about in Platonistic set-theoretical terms, these results show that such a conception is *not necessary* to secure confidence in the body of mathematical practice.

Moving on to the philosophical issues raised by Platonism in set theory, Penrose is right in identifying Gödel as one of the foremost proponents of this position. However, I think it is fair to say that he has few adherents among philosophers of mathematics. One of the more recent ones is Penelope Maddy (1990), though she felt it necessary to develop a compromise form of Platonism (between that of Gödel and Quine); even so that has met with little support. Admittedly, every over-all philosophy of mathematics has its difficulties, but Penrose make it seem that the Platonistic position is a mat-

ter of common consensus, which is not the case for those who have given these questions more than token attention. While one may well agree that questions of truth in the natural numbers are of a determinate character, already the assumption of a supposed definite totality of arbitrary sets of natural numbers is highly problematic. Indeed, Gödel himself, at least for a period in the 1930s, found this deeply troubling. In a previously unpublished lecture (**19330* in Gödel (1995)), he said that: "The result of the preceding discussion is that our axioms [for set theory], if interpreted as meaningful statements, necessarily presuppose a kind of Platonism, which cannot satisfy any critical mind and which does not even produce the conviction that they are consistent." (op.cit. p. 50). And Gödel continued to take proof-theoretical approaches to consistency seriously throughout his life (cf. also **1938a* in Gödel (1995) and the introductory notes to that and **19330*). Incidentally, on p. 116 of *SM*, Penrose says that Paul Cohen, in the last section of his 1966 book on the independence of AC and CH from ZF set theory "reveals himself to be, like Gödel [and Penrose] a true Platonist for whom matters of mathematical truth are *absolute* and not arbitrary." While that is a reasonable inference from what Cohen said there, shortly after that, at a 1967 conference, he stated: "By now it may have become obvious that I have chosen the Formalist position [as opposed to the Platonic Realist position for set theory" (Cohen 1971, p. 13). As far as I know, that is still his position.

Penrose reports in sec. 3.1 on what Gödel took the significance of his incompleteness theorems to be, via a quotation which had circulated some time back from Gödel's unpublished Gibbs lecture of 1951. That piece is now available in full as **1951* in Gödel (1995), with an illuminating introductory note by George Boolos. More cautious than Penrose, Gödel there comes to the conclusion that "*either...the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems.*" (op.cit., p. 310). Boolos' discussion of this is tonic: "There is a gap between the proposition that no finite machine meeting certain weak conditions can print a certain formal sentence (which will depend on the machine) and the statement that if the human mind is a finite machine, there exist truths that cannot be established by any proof the human mind can conceive. ...it is certainly not obvious what it means to say that the human mind, or even the mind of some one human being *is* a finite machine, e.g. a Turing machine. And to say that the mind (at least in its theorem-proving aspect), or *a* mind, may

be represented by a Turing machine is to leave entirely open just *how* it is so represented." (Boolos (1995) p. 293). The same applies *mutatis mutandis* to Penrose's Gödelian argument, and with that, enough said for now.

References

A.R. Anderson (ed.), *Minds and Machines*, Prentice-Hall (N.J.) 1964.

G. Boolos, Introductory note to **1951*, in Gödel (1995), 290–304.

P. Cohen, *Set Theory and the Continuum Hypothesis*, W.A. Benjamin, Inc. (N.Y.), 1966.

———, Remarks on the foundation of set theory, in *Axiomatic Set Theory*, Proc. Symp. Pure Math. XIII, Part I, Amer. Math. Soc. (Providence), 1971.

M. Davis, How subtle is Gödel's theorem? More on Roger Penrose, *Behavioral and Brain Sciences* 16 (1993), 611–612.

S. Feferman, Transfinite recursive progressions of axiomatic theories, *J. Symbolic Logic* 27 (1962), 383–390.

———, Turing in the land of O(z), in *The Universal Turing Machine. A Half-century Survey*, Oxford University Press (Oxford), 1988, 113–147.

———, Hilbert's program relativized: proof-theoretical and foundational reductions, *J. Symbolic Logic* 53, 1988, 364–384.

K. Gödel, *Collected Works, Vol. I. Publications 1929– 1936*, Oxford University Press (N.Y.), 1986.

———, *Collected Works, Vol. II. Publications 1938–1974*, Oxford University Press (N.Y.), 1990.

———, *Collected Works, Vol. III. Unpublished Essays and Lectures*, Oxford University Press (N.Y.), 1995.

S.C. Kleene, *Introduction to Metamathematics*, D. van Nostrand Co. (N.Y.), 1952.

P. Maddy, *Realism in Mathematics*, Oxford University Press (N.Y.), 1990.

R. Penrose, *The Emperor's New Mind*: *Concerning computers, minds and the laws of physics*, Oxford University Press (Oxford), 1989.

———, *Shadows of the Mind*: *A search for the missing science of consciousness*, Oxford University Press (Oxford), 1994.

G. Pólya, *Mathematical Discovery* (Two Vols.), Wiley (N.Y.), 1962.

C. Smorynski, The incompleteness theorems, in *Handbook of Mathematical Logic*, North-Holland Pub. Co. (Amsterdam), 1977, 821–865.

A. Turing, Systems of logic based on ordinals, *Proc. London Math. Soc.* (2) 45 (1939) 161–228.