Gödel's incompleteness theorems, free will and mathematical thought

Solomon Feferman

In memory of Torkel Franzén

Abstract. Some have claimed that Gödel's incompleteness theorems on the formal axiomatic model of mathematical thought can be used to demonstrate that mind is not mechanical, in opposition to a Formalist-Mechanist Thesis. Following an explanation of the incompleteness theorems and their relationship with Turing machines, we here concentrate on the arguments of Gödel (with some caveats) and Lucas among others for such claims; in addition, Lucas brings out the relevance to the free will debate. Both arguments are subject to a number of critiques. The article concludes with the formulation of a modified Formalist-Mechanist Thesis which *prima facie* guarantees partial freedom of the will in the development of mathematical thought.

1. Logic, determinism and free will. The determinism-free will debate is perhaps as old as philosophy itself and has been engaged in from a great variety of points of view including those of scientific, theological and logical character; my concern here is to limit attention to two arguments from logic. To begin with, there is an argument in support of determinism that dates back to Aristotle, if not farther. It rests on acceptance of the Law of Excluded Middle, according to which every proposition is either true or false, no matter whether the proposition is about the past, present or future. In particular, the argument goes, whatever one does or does not do in the future is determined in the present by the truth or falsity of the corresponding proposition. Surely no such argument could really establish determinism, but one is hard pressed to explain where it goes wrong. One now classic dismantling of it has been given by Gilbert Ryle, in the chapter 'What was to be' of his fine book, *Dilemmas* (Ryle 1954). We leave it to the interested reader to pursue that and the subsequent literature.

The second argument coming from logic is much more modern and sophisticated; it appeals to Gödel's incompleteness theorems (Gödel 1931) to make the case against determinism and in favor of free will, insofar as that applies to the mathematical potentialities of human beings. The claim more precisely is that as a consequence of the incompleteness theorems, those potentialities cannot be exactly circumscribed by the output of any computing machine even allowing unlimited time and space for its work. Here there are several notable proponents, including Gödel himself (with caveats), J. R. Lucas and Roger Penrose. All of these arguments have been subject to considerable critical analysis; it is my purpose here to give some idea of the nature of the claims and debates, concluding with some new considerations that may be in favor of a partial mechanist account of the mathematical mind. Before getting into all that we must first give some explanation both of Gödel's theorems and of the idealized machines due to Alan Turing which connect the formal systems that are the subject of the incompleteness theorems with mechanism.

2. Gödel's incompleteness theorems. The incompleteness theorems concern formal axiomatic systems for various parts of mathematics. The reader is no doubt familiar with one form or another of Euclid's axioms for geometry. Those were long considered to be a model of rigorous logical reasoning from first principles. However, it came to be recognized in the 19th century that Euclid's presentation had a number of subtle flaws and gaps, and that led to a much more rigorous presentation of an axiomatic foundation for geometry by David Hilbert in 1899. Hilbert was then emerging as one of the foremost mathematicians of the time, a position he was to hold well into the 20th century. Axiom systems had also been proposed in the late 19th century for other mathematical concepts including the arithmetic of the positive integers by Giuseppe Peano and of the real numbers by Richard Dedekind. In the early 20th century further very important axiom systems were provided by Ernst Zermelo for sets and by Alfred North Whitehead and Bertrand Russell for a proposed logical foundation of mathematics in their massive work, *Principia Mathematica.* Hilbert recognized that these various axiom systems when fully formally specified could themselves be the subject of mathematical study, for example concerning questions of their consistency, completeness and mutual independence of their constitutive axioms.

As currently explained, a specification of a formal axiom system S is given by a specification of its underlying formal language L and the axioms and rules of inference of S. To set up the language L we must prescribe its basic symbols and then say which finite sequences of basic symbols constitute meaningful expressions of the language; moreover, that is to be done in a way that can be effectively checked, i.e. by a finite algorithmic procedure. The sentences ("closed formulas") of L are singled out among its meaningful expressions; they are generated in an effective way from its basic relations by means of the logical operations. If A is a sentence of L and a definite interpretation of the basic relations of L is given in some domain of objects D then A is true or false under that interpretation. The axioms of S are sentences of L and the rules of inference lead from such sentences to new sentences; again, we need to specify these in a way that can be effectively checked. A sentence of L is said to be *provable in* S if it is the last sentence in a *proof from* S, i.e. a finite sequence of sentences each of which is either an axiom or follows from earlier sentences in the sequence by one of the rules of inference. S is *consistent* if there is no sentence A of L such that both A and its negation (not-A) are provable in S. One of the consequences of Gödel's theorems is that there are formal systems S in the language of arithmetic for which S is consistent yet S proves some sentence A which is false in the domain D of positive integers (1, 2, 3,…).

Hilbert introduced the term *metamathematics* for the mathematical study of formal systems for various branches of mathematics. In particular, he proposed as the main program of metamathematics the task of proving the consistency of successively stronger systems of mathematics such as those mentioned above, beginning with the system PA for Peano's Axioms. In order to avoid circularities, Hilbert's program included the proviso that such consistency proofs were to be carried out by the most restrictive mathematical means possible, called *finitistic* by him.

In an effort to carry out Hilbert's program for a substantial part of the formal system PM of *Principia Mathematica* Gödel met a problem which he recognized could turn into a fundamental obstacle for the program. He then recognized that that problem was already met with the system PA. This led to Gödel's stunning theorem that one cannot prove its consistency by any means that can be represented formally within PA, assuming the

consistency of PA. In fact, he showed that if S is any formal system which contains PA either directly or via some translation (as is the case with the theory of sets), and if S is consistent, then the consistency of S cannot be proved by any means that can be carried out within S. This is what is called Gödel's *second incompleteness theorem* or his *theorem on the unprovability of consistency.* The *first incompleteness theorem* was the main way-station to its proof; we take it here in the form that if a formal system S is a consistent extension of PA then there is an arithmetical sentence G which is true but not provable in S, where truth here refers to the standard interpretation of the language of PA in the positive integers. That sentence G (called the Gödel sentence for S) expresses of itself that it is not provable in S.

3. Proofs of the incompleteness theorems. We need to say a bit more about how all this works in order to connect Gödel's theorems with Turing machines. It is not possible to go into full detail about how Gödel's theorems are established, but the interested reader will find that there are now a number of excellent expositions at various levels of accessibility which may be consulted for further elaboration.[1] In order to show for these theorems how various metamathematical notions such as provability, consistency and so on can be expressed in the language of arithmetic, Gödel attached numbers to each symbol in the formal language L of S and then—by using standard techniques for coding finite sequences of numbers by numbers—attached numbers as code to each expression E of L, considered as a finite sequence of basic symbols. These are now called the *Gödel number* of the expression E. In particular, each sentence A of L has a Gödel number. Proofs in S are finite sequences of sentences, and so they too can be given Gödel numbers. Gödel then showed that the Proof-in-S relation, "n is the number of a proof of the sentence with Gödel number m in S", is definable in the language of arithmetic. Hence if A is a sentence of S and m is its Gödel number then the sentence which says there exists an n such that the Proof-in-S relation holds between n and m expresses that A is provable from S. So we can also express directly from this that A is *not* provable from

---

[1] In particular, I would recommend Franzén (2004) for an introduction at a general level, and Franzén (2005) and Smith (2007) for readers with some background in higher mathematics.

S. Next, Gödel used an adaptation of what is called *the diagonal method* to construct a specific sentence G, such that PA proves G is equivalent to the sentence expressing that G is not provable in S. Finally, he showed:

(*) If S is consistent then G is (indeed) not provable from S.

It should be clear from the preceding that the statement that S is consistent, i.e. that there is no A such that both A and not-A are provable in S, can also be expressed in the language of arithmetic; we use Con(S) to denote this statement.

The second incompleteness theorem (unprovability of consistency). If S is a formal system such that S includes PA, and S is consistent, then the sentence Con(S) expressing the consistency of S in arithmetic is not provable in S.

The way Gödel established this is by formalizing the entire argument leading to (*) in Peano Arithmetic. And since the sentence expressing that G is not provable in S is equivalent to G itself, it follows that PA proves:

(**) If Con(S) then G.

So if S were to prove its own consistency statement Con(S) it would also prove G, contrary to (*).

Gödel obtained these remarkable theorems at age 24 as a graduate student at the University of Vienna. The significance of the second incompleteness theorem for Hilbert's program is that if S is a consistent system in which all finitistic methods can be formalized then one cannot give a finitistic consistency proof of S. It was conjectured by Johan von Neumann that all finitistic methods can be formalized in PA and hence that Hilbert's program would already meet a fundamental obstacle at that point. Gödel did not accept von Neumann's conjecture at first but came around to it within a few years and that is now the common point of view. On the other hand, Hilbert apparently never accepted that Gödel's theorem doomed his consistency program to failure.

4. Turing machines and formal systems. Early in the 1930s, two proposals were made by the logicians Alonzo Church and Jacques Herbrand, respectively, to define the concept

of *effective computation procedure* in precise mathematical terms. Gödel found a defect in Herbrand's definition and then modified it so as to avoid its problem. It was then shown by Church and his students that his definition and that of Herbrand-Gödel lead to the same class of computable functions; even so, Gödel did not find either proposal convincing. A couple of years later, the young Cambridge mathematician Alan Turing came up with still another definition in terms of computability on machines of an idealized kind, since then called *Turing machines*. In his paper Turing (1937) also showed the equivalence of his computability notion with those of Church and Herbrand-Gödel. Church quickly accepted Turing's explication of the informal notion of effective computation procedure as being the most convincing of the three then on offer. Gödel apparently did so too, but the first statement by him in print to that effect was not made until almost thirty years later. That was in a postscript he added to the reprinting of lectures that he had given in Princeton 1934 in the collection *The Undecidable* (Davis 1965):

> Turing's work gives an analysis of the concept of "mechanical procedure" (alias "algorithm" or "computation procedure" or "finite combinatorial procedure"). This concept is shown to be equivalent with that of a "Turing machine". *A formal system can simply be defined to be any mechanical procedure for producing formulas, called provable formulas*. For any formal system in this sense there exists one in the [usual] sense … that has the same provable formulas (and likewise vice versa) ... . (Gödel 1965) in (Gödel 1986, p. 369) [Italics mine]

Turing's idea was to isolate the most primitive steps of what human computors actually do. The computational work of following a finite set of rules is that of entering (or erasing) a specified list of symbols in various locations and moving from one location to the next. The amount of space and time needed for carrying out a given computation cannot be fixed in advance. Turing reduced this to working on a (potentially) infinite tape divided into a series of squares, of which at any stage of the computation only a finite number are marked. At any active stage in the computation procedure, exactly one instruction is being followed and exactly one square is being scanned; it may be empty or be marked with one of the symbols. The possible actions for a given instruction, noting

the state of the scanned square, are to enter a specified symbol or erase its contents, move right or left and proceed to another instruction. (If one is at the left end of the tape, the instruction to move left has no effect.) Beginning with any initial configuration starting at the leftmost square the computation terminates—if at all—when one arrives at an instruction that is designated as the final one. A Turing machine M is specified by its instruction set.

The most primitive alphabet for such computations consists of one symbol, the tally |; each positive integer is then represented by a finite sequence of tallies ||…|, successively marked off on the tape, directly preceded and followed by empty squares. To compute a function f of positive integers such as squaring, i.e. $f(n) = n^2$, one enters n tallies as the initial configuration; the computation is to terminate when it is scanning the rightmost tally of a sequence of n^2 tallies. In general, a function f from positive integers to positive integers is said to be *effectively computable by a Turing machine M* if for each input n as initial configuration the procedure terminates with f(n) as output. By an *effectively enumerable set* of positive integers is meant the range {f(1), f(2), f(n),..} of an effectively computable function; there may be repetitions in this range so that it is in fact a finite set. A set is *effectively decidable* if the function f(n) =1 if n is a member of the set and f(n) =2 if it is not (called the characteristic function of the set) is effectively computable. Every non-empty effectively decidable set is effectively enumerable, but Turing showed there are effectively enumerable sets that are not effectively decidable. The notion of effectively computable function is extended in a direct way to those with two or more arguments; a relation between two or more arguments is then effectively decidable if its characteristic function is computable.

Given these definitions, Gödel's above stated identification of the most general notion of formal system with "mechanical procedures for producing formulas" may be spelled out as follows. First of all, given a formal system S, one replaces each formula of the language of S by its Gödel number. By the effectiveness conditions on the specification of S, the set of axioms of S form an effectively decidable set, and each rule of inference, considered as a relation between one or more hypotheses and a conclusion, is an effectively decidable relation between the formulas in each place. It is then an exercise to

show that the Proof-in-S relation is effectively decidable. Now define a function f as follows: if n codes a finite sequence whose last term is m, and n is in the Proof-in-S relation to m, then f(n) = m; otherwise, f(n) is the number of some fixed provable sentence selected in advance. Thus the range of f is exactly the set of (Gödel numbers of) provable formulas of S. In terms of the quote from Gödel above, this shows that *given any formal system there is an associated mechanical procedure for producing its provable formulas*.

Conversely, given a formal language L and a Turing machine M for computing some function f, form the set {f(1), f(2), f(3),…} and then successively eliminate all terms that are not numbers of sentences in L; the result is still effectively enumerable, and its set of purely logical consequences is the set of provable formulas of a suitable formal system S in the language L. Thus *any mechanical procedure may be effectively transformed into another such procedure for producing the provable formulas of a formal system.*

Given Gödel's identification in these senses of formal systems with mechanical procedures, one is led to the following formulation of the thesis that the mathematical mind is mechanical:

The Formalist-Mechanist Thesis I. *Insofar as human mathematical thought is concerned, mind is mechanical in that the set of all mathematical theorems, actual or potential, is the set of provable sentences of some formal system.*

Note well that this is a thesis concerning mathematical thought only. Of course that would be a consequence of all mental activity being determined in some way by a machine. But the thesis is compatible with thought in general not being describable in mechanistic terms. We shall abbreviate this thesis as FMT I.

In the following we shall concentrate on two thinkers who deny FMT I to some extent or other, namely Gödel and Lucas.

5. Gödel on minds and machines. Gödel first laid out his thoughts in this direction in what is usually referred to as his 1951 Gibbs lecture, 'Some basic theorems on the

foundations of mathematics and their implications' (Gödel 1951).[2] The text of this lecture was never published in his lifetime, though he wrote of his intention to do so soon after delivering it. After Gödel died, it languished with a number of other important essays and lectures in his *Nachlass* until it was retrieved for publication in Volume III of Gödel's *Collected Works* (1995, pp. 304-323).

There are essentially two parts to the Gibbs lecture, both drawing conclusions from the incompleteness theorems. The first part concerns the potentialities of mind vs. machines for the discovery of mathematical truths. The second part is an argument aimed to "disprove the view that mathematics is only our own creation", and thus to support some version of platonic realism in mathematics; only the first part concerns us here.[3] Gödel there highlighted the following dichotomy:

> *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 ...* (Gödel 1995, p. 310) [Italics Gödel's]

By a *diophantine problem* is meant a proposition of the language of Peano Arithmetic of a relatively simple form whose truth or falsity is to be determined; its exact description is not important to us.[4] Gödel showed that the consistency of a formal system is equivalent to a diophantine problem, to begin with by expressing it in the form that no number codes a proof of a contradiction. According to Gödel, his dichotomy is a "mathematically

---

[2] Gödel's lecture was the twenty-fifth in a distinguished series set up by the American Mathematical Society to honor the 19th century American mathematician, Josiah Willard Gibbs, famous for his contributions to both pure and applied mathematics. It was delivered to a meeting of the AMS held at Brown University on December 26, 1951.

[3] George Boolos wrote a very useful introductory note to both parts of the Gibbs lecture in Vol. III of the Gödel *Works*. More recently I have published an extensive critical analysis of the first part, under the title "Are there absolutely unsolvable problems? Gödel's dichotomy" (Feferman 2006), followed by the closely related "Gödel, Nagel, minds and machines" (Feferman 2009) on both of which I draw extensively in the following.

[4] Nowadays, mathematicians reserve the terminology 'diophantine equations' and 'diophantine problems' to a more specialized class than taken by Gödel. However, Gödel's have been shown to be equivalent to the non-existence of solutions to suitable diophantine equations.

established fact" which is a consequence of the incompleteness theorem. However, all that he says by way of an argument for it is the following:

> [I]f the human mind were equivalent to a finite machine then objective mathematics not only would be incompletable in the sense of not being contained in any well-defined axiomatic system, but moreover there would exist *absolutely* unsolvable problems…, where the epithet "absolutely" means that they would be undecidable, not just within some particular axiomatic system, but by *any* mathematical proof the mind can conceive. (ibid.) [Italics Gödel's]

By a *finite machine* here Gödel means a Turing machine, and by a *well-defined axiomatic system* he means an effectively specified formal system; as explained above, he takes these to be equivalent in the sense that the set of theorems provable in such a system is the same as the set of theorems that can be effectively enumerated by such a machine. Thus, to say that the human mind is equivalent to a finite machine "even within the realm of pure mathematics" is another way of saying that what the human mind can *in principle* demonstrate in mathematics is the same as the set of theorems of some formal system, i.e. that FMT I holds. By *objective mathematics* Gödel means the totality of true statements of mathematics, which includes the totality of true statements of arithmetic. Then the assertion that objective mathematics is incompletable is simply a consequence of the second incompleteness theorem.

Examined more closely, Gödel's argument is that if the human mind were equivalent to a finite machine, or—what comes to the same thing—an effectively presented formal system S, then there would be *a true statement that could never be humanly proved*, namely Con(S). So that statement would be *absolutely undecidable* by the human mind, and moreover it would be equivalent to a diophantine statement. Note however, the tacit assumption that the human mind is consistent; otherwise, it is equivalent to a formal system in a trivial way, namely one that proves all statements. Actually, Gödel apparently accepts a much stronger assumption, namely that we prove *only* true statements; but for his argument, only the weaker assumption is necessary (together of course with the assumption that PA or some comparable basic system of arithmetic to which the second incompleteness theorem applies has been humanly accepted).

Though he took care to formulate the possibility that the second term of the disjunction holds, there's a lot of evidence outside of the Gibbs lecture that Gödel was convinced of the anti-mechanist position as expressed in the first disjunct. That's supplied, for example, in his informal communication of various ideas about minds and machines to Hao Wang, initially in the book, *From Mathematics to Philosophy* (Wang 1974, pp. 324-326), and then at greater length in *A Logical Journey. From Gödel to Philosophy* (Wang 1996, especially Ch. 6). So why didn't Gödel state that outright in the Gibbs lecture instead of the more cautious disjunction in the dichotomy? The reason was simply that he did not have an unassailable proof of the falsity of the mechanist position. Indeed, despite his views concerning the "impossibility of physico-chemical explanations of … human reason" he raised some caveats in a series of three footnotes to the Gibbs lecture, the second of which is as follows:

> [I]t is conceivable ... that brain physiology would advance so far that it would be known with empirical certainty
>
> 1. that the brain suffices for the explanation of all mental phenomena and is a machine in the sense of Turing;
>
> 2. that such and such is the precise anatomical structure and physiological functioning of the part of the brain which performs mathematical thinking. (ibid.)

Some twenty years later, Georg Kreisel made a similar point in terms of formal systems rather than Turing machines:

> [I]t has been clear since Gödel's discovery of the incompleteness of formal systems that we could not have *mathematical* evidence for the adequacy of any formal system; but this does not refute the possibility that some quite specific system … encompasses all possibilities of (correct) mathematical reasoning …
>
> *In fact the possibility is to be considered that we have some kind of nonmathematical evidence for the adequacy of such [a system]*. (Kreisel 1972, p. 322) [Italics mine]

I shall call the genuine possibility entertained by Gödel and Kreisel, *the mechanist's empirical defense* (or *escape hatch*) against claims to have *proved* that mind exceeds mechanism on the basis of the incompleteness theorems, that is that FMT I is wrong.

6. Lucas on minds and machines. The first outright such claim was made by the Oxford philosopher J. R. Lucas in his article, 'Minds, machines and Gödel' (Lucas 1961): "Gödel's theorem seems to me to prove that Mechanism is false, that is, that minds cannot be explained as machines" (p. 112). His argument is to suppose that there is a candidate machine M (called by him a "cybernetical machine") that enumerates exactly the mathematical sentences that can be established to be true by the human mind, hence exactly what can be proved in a formal system for humanly provable truths. Assuming that,

> [we] now construct a Gödelian formula [the sentence G described in sec. 3 above] in this formal system. This formula cannot be *proved-in-the-system*. Therefore the machine cannot produce the corresponding formula as being true. But we can see that the Gödelian formula is true: any rational being could follow Gödel's argument, and convince himself that the Gödelian formula, although unprovable-in-the-system, was nonetheless…true. … 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. … therefore we cannot hope ever to produce a machine that will be able to do all that a mind can do: we can never not even in principle, have a mechanical model of the mind. (Lucas 1961, p. 115) [Italics Lucas's]

Paul Benacerraf and Hilary Putnam soon objected to Lucas' argument on the grounds that he was assuming it is known that one's mind is consistent, since Gödel's theorem only applies to consistent formal systems. But Lucas had already addressed this as follows:

> … a mind, *if it were really a machine*, could not reach the conclusion that it was a consistent one. [But] for a mind which is not a machine no such conclusion follows. … It therefore seems to me both proper and reasonable for a mind to assert its own consistency: proper, because although machines, as we might have

> expected, are unable to reflect fully on their own performance and powers, yet to be able to be self-conscious in this way is just what we expect of minds: and reasonable, for the reasons given. Not only can we fairly say simply that we *know* we are consistent, apart from our mistakes, but we must in any case *assume* that we are, if thought is to be possible at all; … and finally we can, in a sense, *decide* to be consistent, in the sense that we can resolve not to tolerate inconsistencies in our thinking and speaking, and to eliminate them, if ever they should appear, by withdrawing and cancelling one limb of the contradiction. (ibid., p. 124) [Italics Lucas's]

In this last, there is a whiff of the assertion of human free will. Lucas is more explicit about the connection in the conclusion to his essay:

> If the proof of the falsity of mechanism is valid, it is of the greatest consequence for the whole of philosophy. Since the time of Newton, the bogey of mechanist determinism has obsessed philosophers. If we were to be scientific, it seemed that we must look on human beings as determined automata, and not as autonomous moral agents … But now, though many arguments against human freedom still remain, the argument from mechanism, perhaps the most compelling argument of them all, has lost its power. No longer on this count will it be incumbent on the natural philosopher to deny freedom in the name of science: no longer will the moralist feel the urge to abolish knowledge to make room for faith. We can even begin to see how there could be room for morality, without its being necessary to abolish or even to circumscribe the province of science. Our argument has set no limits to scientific enquiry: it will still be possible to investigate the working of the brain. It will still be possible to produce mechanical models of the mind. Only, now we can see that no mechanical model will be completely adequate, nor any explanations in purely mechanist terms. We can produce models and explanations, and they will be illuminating: but, however far they go, there will always remain more to be said. There is no arbitrary bound to scientific enquiry: but no scientific enquiry can ever exhaust the infinite variety of the human mind. (ibid., p. 127)

According to Lucas, then, FMT I is *in principle* false, though there can be scientific evidence for the mechanical workings of the mind to some extent or other insofar as mathematics is concerned. What his arguments do not countenance is the possibility of obtaining fully convincing empirical support for the mechanist thesis, namely that eventually all evidence points to mind being mechanical though we cannot ever hope to supply a *complete perfect description* of a formal system which accounts for its workings.[5] Moreover, such a putative system need not necessarily be consistent. Without such a perfect description for a consistent system as a model of the mind, the argument for Gödel's theorem cannot apply. Lucas, in response to such a suggestion has tried to shift the burden to the mechanist: "The consistency of the machine is established not by the mathematical ability of the mind, but on the word of the mechanist" (Lucas 1996), a burden that the mechanist can refuse to shoulder by simply citing this empirical defense. Finally, the compatibility of FMT I with a non-mechanistic account for thought in general would still leave an enormous amount of room for morality and the exercise of free will.

Despite such criticisms, Lucas has stoutly defended to the present day his case against the mechanist on Gödelian grounds. One can find on his home page[6] most of his published rejoinders to various of these as well as further useful references to the debate. The above quotations do not by any means exhaust the claims and arguments in his thoroughly thought out discussions.

7. Critiques of Gödelian arguments against mechanism. [7] Roger Penrose is another well-known defender of the Gödelian basis for anti-mechanism, most notably in his two books, *The Emperor's New Mind* (1989), and *Shadows of the Mind* (1994). Sensitive to the objections to Lucas, he claimed in the latter only to have proved something more modest (and in accord with experience) from the incompleteness theorems: "Human mathematicians are not using a knowably sound algorithm in order to ascertain

---

[5] That would be analogous to obtaining fully convincing empirical support for the thesis that the workings of, say, the human auditory and visual systems are fully explicable in neurological and physical terms, though one will never be able to produce a complete perfect description of how those operate. I presume that we are in fact in such a position.

[6] http://users.ox.ac.uk/~jrlucas/

[7] This section is drawn directly from (Feferman 2009).

mathematical truth." (Penrose 1994, p. 76). But later in that work he came up with a new argument purported to show that the human mathematician can't even consistently *believe* that his mathematical thought is circumscribed by a mechanical algorithm (Penrose 1994, secs. 3.16 and 3.23). Extensive critiques have been made of Penrose's original and new arguments in an issue of the journal PSYCHE (1996), to which he responded in the same issue. And more recently, Stewart Shapiro (2003) and Per Lindström (2001, 2006) have carefully analyzed and critiqued his "new argument." But Penrose has continued to defend it, as he did in his public lecture for the Gödel Centenary Conference held in Vienna in April 2006.

Historically, there are many examples of mathematical proofs of what can't be done in mathematics by specific procedures, e.g. the squaring of the circle, or the solution by radicals of the quintic, or the solvability of the halting problem. But it is hubris to think that by mathematics alone we can determine what the mathematician can or cannot do in general. The claims by Gödel, Lucas and Penrose to do just that from the incompleteness theorems depend on making highly idealized assumptions both about the nature of mind and the nature of machines. A very useful critical examination of these claims and the underlying assumptions has been made by Shapiro in his article, 'Incompleteness, mechanism and optimism' (1998), among which are the following. First of all, how are we to understand the mathematizing capacity of the human mind, since what is at issue is the producibility of an infinite set of propositions? No one mathematician, whose life is finitely limited, can produce such a list, so either what one is talking about is what the individual mathematician *could do in principle*, or we are talking in some sense about the potentialities of the pooled efforts of the community of mathematicians now or ever to exist. But even that must be regarded as a matter of what can be done *in principle,* since it is most likely that the human race will eventually be wiped out either by natural causes or through its own self-destructive tendencies by the time the sun ceases to support life on earth.

What about the assumption that the human mind is consistent? In practice, mathematicians certainly make errors and thence arrive at false conclusions that in some cases go long undetected. Penrose, among others, has pointed out that when errors are

detected, mathematicians seek out their source and correct them (cf. Penrose 1996, pp. 137 ff), and so he has argued that it is reasonable to ascribe self-correctability and hence consistency to our idealized mathematician. But even if such a one can correct all his errors, can he know with mathematical certitude, as required for Gödel's claim, that he is consistent?

As Shapiro points out, the relation of both of these idealizations to practice is analogous to the competence/performance distinction in linguistics.

There are two further points of idealization to be added to those considered by Shapiro. The first of these is the assumption that the notions and statements of mathematics are fully and faithfully expressible in a formal language, so that what can be humanly proved can be compared with what can be the output of a machine. In this respect it is usually pointed out that the only part of the assumption that needs be made is that the notions and statements of elementary number theory are fully and faithfully represented in the language of first-order arithmetic, and that among those only simply universal ("diophantine") statements need be considered, since that is the arithmetized form of the consistency statements for formal systems. But even this idealization requires that statements of unlimited size must be accessible to human comprehension.

Finally to be questioned is the identification of the notion of finite machine with that of Turing machine. Turing's widely accepted explication of the informal concept of effective computability puts no restriction on time or space that might be required to carry out computations. But the point of that idealization was to give the strongest *negative* results, to show that certain kinds of problems can't be decided by a computing machine, no matter how much time and space we allow. And so if we carry the Turing analysis over to the potentiality of mind in its mathematizing capacity, to say that mind infinitely surpasses any finite machine is to say something even stronger. It would be truly impressive if that could be definitively established, but none of the arguments that have been offered are resistant to the mechanist's empirical defense. Moreover, suppose that the mechanist is right, and that in some reasonable sense mind *is* equivalent to a finite machine; is it appropriate to formulate that in terms of the identification of what is humanly provable with what can be enumerated by a Turing machine? Isn't the

mechanist aiming at something stronger in the opposite direction, namely an explanation of the mechanisms that govern the production of human proofs?

8. Mechanism and partial freedom of the will. This last point is where I think something new has to be said, something that I already drew attention to in (Feferman 2006, 2009). Namely, there is an *equivocation* involved, that lies in identifying *how* the mathematical mind works with the totality of *what* it can prove. Again, the difference is analogous to what is met in the study of natural language, where we are concerned with the *way* in which linguistically correct utterances are generated and *not* with the potential totality of *all* such utterances. That would seem to suggest that if one is to consider *any* idealized formulation of the mechanist's position at all in logical terms, it ought to be of the mind as one *constrained* by the axioms and rules of some effectively presented formal system. Since in following those axioms and rules one has *choices* to be made at each step, *at best* that identifies the mathematizing mind with *the program for a non-deterministic Turing machine*, and *not* with the set of its enumerable statements (even though that can equally well be supplied by a deterministic Turing machine).[8] One could no more disprove this modified version of the idealized mechanist's thesis than the version considered by Gödel, et al., simply by applying the mechanist's empiricist argument. Nevertheless, it is difficult to conceive of any formal system of the sort with which we are familiar, from Peano Arithmetic (PA) up to Zermelo-Fraenkel Set Theory (ZF) and beyond, actually underlying mathematical thought as it is experienced.

As I see it, a principal reason for the implausibility of this modified version of the mechanist's thesis lies in the concept of a formal system S that is currently taken for granted in logical work. An essential part of that concept is that the language L of S is fixed once and for all. For example, the language of PA is determined (in one version) by taking the basic symbols to be those for equality, zero, successor, addition and multiplication and that of ZF is fixed by taking its basic symbols to be those for equality and membership. This forces axiom schemata that may be used in such systems, such as for mathematical induction in arithmetic and separation in set theory, to be infinite

---

[8] Lucas (1961, pp. 113-114), recognized the equivalence of non-deterministic and deterministic Turing machines with respect to the set of theorems proved by each.

bundles of all possible substitution instances by formulas from that language; this makes metamathematical but not mathematical sense. Besides that, the restriction of mathematical discourse to a language fixed in advance, even if only implicitly, is completely foreign to mathematical practice.

In recent years I have undertaken the development of a modified conception of formal system that does justice to the openness of practice and yet gives it an underlying rule-governed logical-axiomatic structure; it thus suggests a way, admittedly rather speculative, of straddling the Gödelian dichotomy. This is in terms of a notion of *open-ended schematic axiomatic system*, i.e. one whose schemata are finitely specified by means of propositional and predicate variables (thus putting the 'form' back into 'formal systems') while the language of such a system is considered to be *open-ended,* in the sense that its basic vocabulary may be expanded to any wider conceptual context in which its notions and axioms may be appropriately applied. In other words, on this approach, *implicit in the acceptance of given schemata is the acceptance of any meaningful substitution instances that one may come to meet,* but which those instances are is not determined by restriction to a specific language fixed in advance (cf. Feferman 1996 and 2006a, and Feferman and Strahm 2000). The idea is familiar from logic with such basic principles as "P & Q implies P" and rules such as, "from P and P implies Q, infer Q", for arbitrary propositions P and Q. But it is directly extended to the principle of mathematical induction for any property P ("if P(1) and for all n, P(n) implies P(n+1), then for all positive integers n, P(n)"), and Zermelo's separation axiom for any property P ("if a is any set then there is a set b such that for all x, x is a member of b if and only if x is a member of a and P(x) holds"). All of these may be considered (and are actually employed) without restriction to any specific language fixed in advance.

This leads me to suggest the following revision of FMT I:

The Formalist-Mechanist Thesis II. *Insofar as human mathematical thought is concerned, mind is mechanical in that it is completely constrained by some open-ended schematic formal system.*

If the concepts of mathematics turned out to be limited to those that can be expressed in one basic formal language L, the two theses would be equivalent. So the point of this second thesis is that the conceptual vocabulary of mathematics is not necessarily limited in that way, but that mathematics is otherwise constrained once and for all by the claimed finite number of open-ended schematic principles and rules. The idea is spelled out in the final section of (Feferman 2009), to which the reader is referred given the limitations of space here. But I will repeat some of the arguments as to why the language of mathematics should be considered to be open-ended, i.e. not restricted to one language L once and for all.

Consider, to the contrary, the claim by many that all mathematical concepts are definable in the language of axiomatic set theory. It is indeed the case that the current concepts of working ("pure") mathematicians are with few exceptions all expressible in set theory. But there are genuine outliers. For example a natural and to all appearances coherent mathematical notion whose full use is not set-theoretically definable is that of a category; only so-called "small" categories can be directly treated in that way (cf. Mac Lane 1971 and Feferman 1977 and 2006b). Other outliers are to be found on the constructive fringe of mathematics in the schools of Brouwerian intuitionism and Bishop's constructivism (cf. Beeson 1985) whose basic notions and principles are not directly accounted for in set theory with its essential use of classical logic. And it may be argued that there are informal mathematical concepts like those of knots, or infinitesimal displacements on a smooth surface, or of random variables, to name just a few, which may be the subject of convincing mathematical reasoning but that are accounted for in set theory only by some substitute notions that share the main expected properties but are not explications in the ordinary sense of the word. Moreover, the idea that set-theoretical concepts and questions like Cantor's continuum problem have determinate mathematical meaning has been challenged on philosophical grounds (Feferman 2000). Finally, there is a theoretical argument for openness, even if one accepts the language L of set theory as a determinately meaningful one. Namely, by Tarski's theorem, the notion of truth T_L for L is not definable in L; and then the notion of truth for the language obtained by adjoining T_L to L is not definable in *that* language, and so on (even into the transfinite).

Another argument that may be made against the restriction of mathematics to a language fixed in advance is historical. Simply witness the progressive amplification of the body of mathematical concepts since the emergence of abstract mathematics in Greek times. It would be hubris to suppose that that process will ever be brought to completion. But having generally granted that certain open-ended schematic principles and rules completely govern all *logical* thinking, it is not hubris to grant the there are some finitely many open-ended schematic principles and rules that completely constrain all *mathematical* thinking now and ever to come, no matter what new concepts and specifically associated principles one comes to accept. Of course, by the mechanist's empiricist argument one could no more disprove this version FMT II of the mechanist's thesis than the version FMT I considered by Gödel, Lucas, Penrose, et al.[9]

That mathematics is constrained by its modes of reasoning in some way or other accords with ordinary experience; that and much work in the formalization of mathematical thought is what gives plausibility to the FMT II thesis. That the practice of mathematics provides extraordinary scope for the exercise of creative free will is also a feature of everyday experience. But that that free will may only be partial in the sense of FMT II need be no more surprising than that the human exercise of free will as applied to bodily actions is constrained by the laws of natural science. I am taking all this in a *prima facie* sense. Lacking any sort of convincing argument for genuine free will, what is at issue here is whether the laws of nature or thought as we know them leave open the possibility of making real choices at each step in our physical and intellectual lives, in particular in

---

[9] Another kind of suggested combination of openness with mechanism that could evade the arguments from Gödel's theorems has recently been brought to my attention by Martin Solomon: "[I]f we treat mechanical theorem recognizers as *open* systems, continually interacting with their environment, they may enjoy increases in power (possibly through 'random inspirations' from this environment) which could occur in a 'surprising', i.e. non-computable manner." (Lyngzeidetson and Solomon (1994), p. 552) The idea is that these systems respond to varying input data that may be non-computable in a completely computable way in order to generate mathematical theorems. However, no criterion could be built into such a system to insure that only true statements are proved. My open-ended schematic formal systems are also not immune to that problem if faulty concepts are adopted. For example, the concept of "feasibly computable number" leads to the conclusion that all numbers are feasible by applying the induction scheme.

the case of mathematical thought, new choices as to the concepts with which mathematics may deal.

On the other hand, one may well ask to what extent FMT II fits with a mechanistic view of the human mind as a whole.[10] Indeed, one ought to pose that of the stronger thesis FMT I as well, even though Gödel, Lucas and Penrose all considered that a proof of its falsity would amount to a rejection of mechanism, at least in the mental realm. There are actually two competing mechanistic theories of the mind in current cognitive science, the digital computational model identified with such figures as Alan Turing and John von Neumann, and the connectionist computational model exemplified by the work of John McClelland and David Rumelhart; see (Harnish 2002) for an excellent historical introduction to and expository survey of these two approaches. It is only the digital computational conception that is the target of the anti-mechanists who argue from Gödel's incompleteness theorems. Though FMT I is a consequence of that viewpoint, the converse does not hold since FMT I only concerns the mathematizing capacities of the human mind. Despite the empirical defence in possible support of it, the evidence for FMT I in mathematical practice is actually practically nil. Nevertheless, what has been at issue here is whether it can be disproved on the basis of Gödel's theorems, and I have argued along with others that it cannot. If that is granted, it is theoretically possible that FMT I holds without that making the case for the digital computational model of the mind as a whole. FMT II is even farther from that point of view but it seems to me to be similar enough to FMT I to warrant being called a Formalist-Mechanist Thesis. Speaking for myself, I believe something like FMT II is true but do not subscribe to any mechanistic conception of the mind as a whole.

Stanford University

Email: feferman@stanford.edu

References

---

[10] I wish to thank Mr Lucas for raising this question.

Beeson, M. (1985), Foundations of Constructive Mathematics (Berlin: Springer-Verlag).

Davis, M. (ed.) (1965), The Undecidable. Basic Papers on Undecidable Propositions, Unsolvable problems and Computable Functions (Hewlett, NY: Raven Press).

Feferman, S. (1977), 'Categorical Foundations and Foundations of Category Theory', in (R.E. Butts and J. Hintikka, eds.) Logic, Foundations of Mathematics and Computability Theory, Vol. I (Dordrecht: Reidel), pp. 149-165.

_____ (1995), 'Penrose's Gödelian Argument', PSYCHE 2, pp. 21-32; also at http://psyche.cs.monash.edu.au/v2/psyche-2-07-feferman.html

_____ (1996), 'Gödel's Program for New Axioms: Why, Where, How and What?', in (P. Hájek, ed.) Gödel '96, Lecture Notes in Logic 6, pp. 3-22.

_____ (2006), 'Are There Absolutely Unsolvable Problems? Gödel's Dichotomy', Philosophia Mathematica, Ser. III, 14, pp. 134-152.

_____ (2006a), 'Open-ended Schematic Axiom Systems' (abstract), Bull. Symbolic Logic 12, p. 145.

_____ (2006b), 'Enriched Stratified Systems for the Foundations of Category Theory', in (G. Sica, ed.) What is Category Theory? (Monza: Polimetrica), pp. 185-203.

Feferman, S. and T. Strahm (2000), 'The Unfolding of Non-finitist Arithmetic', Annals of Pure and Applied Logic 104, pp. 75-96.

Franzén, T. (2004), Inexhaustibility: A Non-exhaustive Treatment (Wellesley, MA: A.K. Peters).

_____ (2005), Gödel's Theorem. An Incomplete Guide to its Use and Abuse (Wellesley, MA: A.K. Peters).

Gödel, K. (1931), 'Über Formal Unentscheidbare Sätze der Principia Mathematica und Verwandter Systeme I', Monatshefte für Mathematik und Physik 38, pp.173-198; reprinted with facing English translation in (Gödel 1986), pp. 144-195.

_____ (1934), 'On Undecidable Propositions of Formal Mathematical Systems', (mimeographed lecture notes at the Institute for Advanced Study, Princeton, NJ); reprinted in (Davis 1965), pp. 39-74 and (Gödel 1986), pp. 346-371.

_____ (1951), 'Some Basic Theorems on the Foundations of Mathematics and Their Implications', in (Gödel 1995), pp. 304-323.

_____ (1986), Collected Works, Vol. I: Publications 1929-1936 (S. Feferman, et al., eds.) (New York: Oxford University Press).

_____ (1995), Collected Works, Vol. III: Unpublished Essays and Lectures (S. Feferman, et al., eds.) (New York: Oxford University Press).

Harnish, R. M. (2002), Minds, Brains and Computers. An Historical Introduction to the Foundations of Cognitive Science (Oxford: Blackwell Publishers).

Kreisel, G. (1972), 'Which Number-theoretic Problems Can be Solved in Recursive Progressions on Π^1_1 Paths Through O?', J. Symbolic Logic 37, pp. 311-334.

Lindström, P. (2001), 'Penrose's New Argument', J. Philosophical Logic 30, pp. 241-250.

_____ (2006), 'Remarks on Penrose's "New Argument" ', J. Philosophical Logic 35, pp. 231-237.

Lucas, J. R. (1961), 'Minds, Machines and Gödel', Philosophy 36, pp. 112-137.

_____ (1996), 'Minds, Machines and Gödel: A Retrospect', in (P. J. R. Millican and A. Clark, eds.), Machines and Thought: The Legacy of Alan Turing, vol. 1 (Oxford: Oxford University Press), pp. 103-124.

Lyngzeidetson, A. E. and M. K. Solomon (1994), 'Abstract Complexity and the Mind-Machine Problem', British Journal for the Philosophy of Science 45, 549-554.

Mac Lane, S. (1971), Categories for the Working Mathematician (Berlin: Springer-Verlag).

Penrose, R.(1989), The Emperor's New Mind (Oxford: Oxford University Press).

_____(1994), Shadows of the Mind (Oxford: Oxford University Press).

_____ (1996), 'Beyond the Doubting of a Shadow', PSYCHE 2, 89-129; also at http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html .

Ryle, G. (1954), Dilemmas (Cambridge: Cambridge University Press).

Shapiro, S. (1998), 'Incompleteness, Mechanism, and Optimism', Bulletin of Symbolic Logic 4, pp. 273-302.

_____ (2003), 'Mechanism, Truth, and Penrose's New Argument', J. Philosophical Logic 32, pp. 19-42.

Smith, P. (2007), An Introduction to Gödel's Theorems (Cambridge: Cambridge University Press).

Turing, A. M. (1937), 'On Computable Numbers, With an Application to the Entscheidungsproblem', Proc. London Math. Soc. (2) 42, pp. 230-265; correction, ibid. 43, pp. 544-546; reprinted in (Davis 1965, pp. 116-154) and in (Turing 2001, pp. 18-56).

_____ (2001), Collected Works of A. M. Turing. Mathematical Logic (R. O. Gandy and C. E. M. Yates, eds.), (Amsterdam: North-Holland/Elsevier).

Wang, H. (1974) From Mathematics to Philosophy (London: Routledge and Kegan Paul).

_____ (1996), A Logical Journey. From Gödel to Philosophy (Cambridge, MA: MIT Press).