GÖDEL, NAGEL, MINDS, AND MACHINES*

Some fifty years ago (1957 to be exact), Ernest Nagel and Kurt Gödel became involved in a contentious exchange about the possible inclusion of Gödel's original work on incompleteness in the book, *Gödel's Proof,* then being written by Nagel with James R. Newman.[1] What led to the conflict were some unprecedented demands that Gödel made over the use of his material and his involvement in the contents of the book—demands that resulted in an explosive reaction on Nagel's part. In the end the proposal came to naught. But the story is of interest because of what was basically at issue, namely their provocative related but contrasting views on the possible significance of Gödel's theorems for minds versus machines in the development of mathematics. That is our point of departure for the attempts by Gödel, and later J.R. Lucas and Roger Penrose, to establish definitive consequences of those theorems, attempts which—as we shall see—depend on highly idealized and problematic assumptions about minds, machines, and mathematics. In particular, I shall argue that there is a fundamental equivocation involved in those assumptions that needs to be reexamined. In conclusion, that will lead us to a new way of looking at how the mind may work in deriving mathematics which straddles the mechanist and anti-mechanist viewpoints.

The story of the conflict between Gödel and Nagel has been told in full in the introductory note by Charles Parsons and Wilfried Sieg to the correspondence between them in Volume V of the Gödel *Collected Works*,[2] so I shall confine myself to the high points.

The first popular exposition of Gödel's incompleteness theorems was published by Nagel and Newman in 1956 in an article entitled "Goedel's Proof" for the *Scientific American*.[3] The article was reprinted soon after in the four volume anthology edited by Newman, *The World of Mathematics: A Small Library of the Literature of Mathematics from A' h-mosé*

* Revised text of the Ernest Nagel Lecture given at Columbia University on September 27, 2007. I wish to thank Hannes Leitgeb and Carol Rovane for their helpful comments on an earlier version.

[1] Nagel and Newman, *Gödel's Proof* (New York: University Press, 1958); revised edition edited by Douglas R. Hofstadter (New York: University Press, 2001).

[2] Gödel, *Collected Works, Volume V: Correspondence H–Z*, Solomon Feferman et al., eds., (New York: Oxford, 2003), pp. 135ff.

[3] *Scientific American*, CXCIV (June 1956): 71–86.

0022-362X/09/0604/201–19

the Scribe to Albert Einstein, Presented with Commentaries and Notes.[4] That was an instant best-seller, and has since been reprinted many times. Newman had been trained as a mathematician but then became a lawyer and was in government service during World War II. Endlessly fascinated with mathematics, he became a member of the editorial board of the *Scientific American* a few years after the war.

Nagel had long been recognized as one of the leading philosophers of science in the United States, along with Rudolf Carnap, Carl Hempel, and Hans Reichenbach. Like them, he was an immigrant to the U.S.A., but unlike them he had come much earlier, in 1911 at the age of ten. Later, while teaching in the public schools, Nagel received his bachelor's degree at City College of New York in 1923 and his Ph.D. in philosophy at Columbia University in 1930. Except for one year at Rockefeller University, his entire academic career was spent at Columbia, where he became the John Dewey Professor of Philosophy and was eventually appointed to the prestigious rank of University Professor in 1967. In his philosophical work, Nagel combined the viewpoints of logical positivism and pragmatic naturalism. His teacher at City College had been Morris R. Cohen, and with Cohen in 1934 he published *An Introduction to Logic and Scientific Method*,[5] one of the first and most successful textbooks in those subjects.

Soon after the appearance of Nagel and Newman's article on Gödel's theorems in *The World of Mathematics*, they undertook to expand it to a short book to be published by New York University Press. Moreover, they had the idea to add an appendix which would include a translation of Gödel's 1931 paper on undecidable propositions together with the notes for lectures on that work that he had given during his first visit to Princeton in 1934.[6] Early in 1957 their editor at the Press, Allan Angoff, approached Gödel for permission to use that material. Though Gödel said he liked the Nagel-Newman article very much as a nontechnical introduction to his work, he said that he was concerned with some troublesome mistakes[7] that had been made in it

[4] In *The World of Mathematics: A Small Library of the Literature of Mathematics from A'h-mosé the Scribe to Albert Einstein, Presented with Commentaries and Notes*, Volume 3 (New York: Simon and Schuster, 1956), pp. 1668–95.

[5] Nagel and Cohen, *An Introduction to Logic and Scientific Method* (New York: Harcourt Brace, 1934).

[6] Gödel, "Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I," *Monatshefte für Mathematik und Physik*, XXXVIII (1931): 173–98; reprinted with facing English translation in Gödel, *Collected Works, Volume I: Publications 1929–1936*, Feferman et al., eds. (New York: Oxford, 1986), pp. 144–95.

[7] See Hilary Putnam, "Review of Nagel and Newman (1958)," *Philosophy of Science* XXVII, 2 (1960): 205–07, for a review of Nagel and Newman's *Gödel's Proof* in which

and even more so with the interpretation of his results, so he was reluctant about agreeing to the proposal. To encourage him to accept, Nagel paid a lengthy personal visit to him at the Institute in March of 1957. A week later, Gödel wrote Angoff with three conditions. First,

> I would have to write an introduction to the appendix on the one hand in order to mention advances that have been made after the publication of my papers, [and] on the other hand in order to supplement the considerations given in the book, about the philosophical implications of my results. I am not very well pleased with the treatment of these questions that came out in the Scientific American and in the "World of Mathematics". This, for the most part is not the fault of the authors, because almost nothing has been published on this subject, while I have been thinking about it in the past few years.[8]

A second condition, somewhat of a surprise, was that he share in some way in the royalties of the book. But the most contentious condition was the one he put at the beginning of his letter to Angoff: "In view of the fact that giving my consent to this plan implies, in some sense, an approval of the book on my part, I would have to see the manuscript and the proof sheets of the book, including the appendix." This condition was made even more explicit in a follow-up letter:

> Of course I shall have to see the manuscript of the book before I sign the contract, so that I can be sure that I am in agreement with its content, or that passages with which I don't agree can be eliminated, or that I may express my view about the questions concerned in the introduction (*ibid.*, p. 4).

Gödel also wrote Nagel saying that he had made to Angoff "the same suggestions I mentioned to you in our conversation in Princeton." But when Nagel was shown the correspondence with Angoff, he exploded. He thought the proposal to share royalties was "grasping and unreasonable" yet was prepared to accept that. But what really ticked him off, as he wrote Angoff, was Gödel's number one condition:

> I could scarcely believe my eyes when I read his ultimatum that he is not only to see the manuscript of our essay *before* signing the contract with you, but that he is to have the right to eliminate anything in the essay of which he disapproves. In short, he stipulates as a condition of signing the contract the right of censorship.

---

several errors are identified, the most egregious being the misstatement of Gödel's first incompleteness theorem and Rosser's improvement thereof on p. 91.

[8] Gödel, *Collected Works, Volume IV: Correspondence A–G*, Feferman et al., eds. (New York: Oxford, 2003), p. 1.

> This seems to me just insulting, and I decline to be a party to any such agreement with Gödel .... If [his] conditions were granted, Jim and I would be compelled to make any alterations Gödel might dictate, and we would be at the mercy of his tastes and procrastination for a period without foreseeable end .... Gödel is of course a great man, but I decline to be his slave.[9]

Nagel's fears about how things might go if they agreed to this condition were indeed well founded, but he left it in Angoff's hands to communicate his displeasure with it. In the event, Angoff tried to avoid direct conflict, and wrote Gödel in a seemingly accomodating but ambiguous way. It was not until August 1957 that Nagel wrote Gödel himself making clear his refusal to accept it:

> ... I must say, quite frankly, that your ... stipulation was a shocking surprise to me, since you were ostensibly asking for the *right to censor* anything of which you disapproved in our essay. Neither Mr. Newman nor I felt we could concur in such a demand without a complete loss of self-respect. I made all this plain to Mr. Angoff when I wrote him last spring, though it seems he never stated our case to you. I regret now that I did not write you myself, for I believe you would have immediately recognized the justice of our demurrer... (*ibid.*, pp. 152–53).

This was passed over in silence in Gödel's response to Nagel. In the end, the proposal to include anything by him in the book by Nagel and Newman came to naught, and even the specific technical errors which Gödel could have brought to their attention remained uncorrected. Since this was to be the first popular exposition that would reach a wide audience, Gödel had good reason to be concerned, but by putting the conditions in the way that he did he passed up the chance to be a constructive critic.

As an aside, it is questionable whether Gödel indeed recognized the "justice of [their] demurrer." He made a similar request seven years later, when Paul Benacerraf and Hilary Putnam approached him about including his papers on Russell's mathematical logic and Cantor's continuum problem in their forthcoming collection of articles on the philosophy of mathematics. Gödel feared that Benacerraf and Putnam would use their introduction to mount an attack on his platonistic views, and he demanded what amounted to editorial control of it as a condition for inclusion. In that case, the editors refused straight off, but gave Gödel sufficient reassurance about the nature of its contents for him to grant permission to reprint as requested.[10]

[9] Gödel, *Collected Works, Volume V: Correspondence H–Z*, pp. 138–39.

[10] Cf. Gödel, *Collected Works, Volume II: Publications 1938–1974*, Feferman et al., eds. (New York: Oxford, 1990), p. 166.

I. GÖDEL'S CONCERNS, PART I: THE FORMULATION OF THE INCOMPLETENESS THEOREMS

Recall that from his first letter to Angoff, one of the conditions that Gödel put for the proposed appendix to the Nagel and Newman book was that he write an introduction to it "in order to mention advances that have been made after the publication of my papers" as well as "to supplement the considerations, given in the book, about the philosophical implications of my results," with which he was "not very well pleased." As to the first of these, Gödel would have wanted to use the opportunity to put on record what he considered the strongest formulation of his incompleteness theorems. This was signaled in a letter that he composed to Nagel early in 1957 but that was apparently never sent. (A number of letters found in Gödel's *Nachlass* were marked *nicht abgeschickt*.) He there writes that

> [c]onsiderable advances have been made ... since 1934. ...it was only by Turing's work that it became completely clear, that my proof is applicable to *every* formal system containing arithmetic. I think the reader has a right to be informed about the present state of affairs.[11]

What he is referring to, of course, is Alan Turing's analysis in 1937 of the concept of effective computation procedure by means of what we now call Turing machines; Gödel had readily embraced Turing's explication after having rejected earlier proposals by Alonzo Church and Jacques Herbrand. But it was not until eight years after the debacle with Nagel and Newman that Gödel spelled out the connection with formal systems and his own work. That was in a postscript he added to the 1965 reprinting of his Princeton lectures in the volume, *The Undecidable*,[12] edited by Martin Davis:

> In consequence of later advances, in particular of the fact that, due to A. M. Turing's work, a precise and unquestionably adequate definition of the general concept of formal system can now be given, the existence of undecidable arithmetical propositions and the non-demonstrability of the consistency of a system in the same system can now be proved rigorously for *every* consistent formal system containing a certain amount of finitary number theory.
>
> Turing's work gives an analysis of the concept of "mechanical procedure" (alias "algorithm" or "computation procedure" or "finite com-

[11] Gödel, *Collected Works, Volume V: Correspondence H–Z*, p. 147.

[12] In Martin Davis, ed., *The Undecidable: Basic Papers on Undecidable Propositions, Unsolvable Problems and Computable Functions* (Hewlett, NY: Raven, 1965); reproduced in Gödel, *Collected Works, Volume I: Publications 1929–1936*, p. 369.

> binatorial 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) ... [Italics mine]

As we will see, there is much more to this identification than meets the eye; in fact, in my view it is the source of a crucial misdirection in the minds versus machines disputes that we shall take up below. By comparison, Nagel and Newman distinguish the consequences of the incompleteness theorems for axiomatic systems and for "calculating machines" in their concluding reflections as follows:

> [Gödel's theorems] show that there is an endless number of true arithmetical statements which cannot be formally deduced from any specified set of axioms .... It follows, therefore, that an axiomatic approach to number theory ... cannot exhaust the domain of arithmetical truth ....[13]

> Gödel's conclusions also have a bearing on the question whether calculating machines can be constructed which would be substitutes for a living mathematical intelligence. Such machines, as currently constructed and planned, operate in obedience to a fixed set of directives built in, and they involve mechanisms which proceed in a step-by-step manner. But in the light of Gödel's incompleteness theorem, there is an endless set of problems in elementary number theory for which such machines are inherently incapable of supplying answers, however complex their built-in mechanisms may be and however rapid their operations (*ibid.*, p. 1695).

Of course, Gödel could charitably read their informal idea of calculating machines simply to be explicated by the notion of Turing machines, but as we saw above, he would have gone beyond that to stress that there is no essential difference between these results.

The first thing that Gödel was surely reacting to on the philosophical rather than the technical side was the statement by Nagel and Newman that

> [the incompleteness theorems] seem to show that the hope of finding an absolute proof of consistency for any deductive system in which the whole of arithmetic is expressible cannot be realized, if such a proof must satisfy the finitistic requirements of Hilbert's original program (*ibid.*, p. 1694).

What must have annoyed Gödel was that this, together with their further reflections on the significance of the incompleteness theorems concerned matters to which he had given a good deal of thought

[13] Nagel and Newman, "Goedel's Proof," *The World of Mathematics*, p. 1694.

over the years, in some respects with overlapping conclusions, in others contrary ones, but in all cases in much greater depth and with much greater care and precision. The trouble was that almost none of his thought on these questions had been published. With respect to the significance of the incompleteness theorems for Hilbert's finitist consistency program this was something that he had made only one brief cautious comment about at the end of his 1931 paper on undecidable propositions. But it was a matter he kept coming back to all through his life, as we found when we unearthed unpublished lectures and seminar presentations in his *Nachlass*. The first time his further views on this would begin to come out in print would be in 1958, a year after the imbroglio with Nagel and Newman.[14] Since my main concern in the remainder of this article is with the minds versus machines debate as it relates to the incompleteness theorems, I shall leave that matter at that.

II. GÖDEL'S CONCERNS, PART II: SIGNIFICANCE OF THE INCOMPLETENESS THEOREMS FOR THE MINDS VERSUS MACHINES DEBATE

The above quotations from Nagel and Newman's 1956 version of "Goedel's Proof," all come from their final three paragraphs on the "far reaching import" of the incompleteness theorems. (My guess is that Nagel was largely responsible for their formulation, but I have no specific evidence for that.) These, with some rewording but no essential change in content, were to form the entire last chapter, entitled "Concluding Reflections," of the 1958 version of the book *Gödel's Proof*; it is only the 1956 version that Gödel would have seen, however, and to which he would have reacted and so it is from there that I go on to draw the quotations. We now continue with the remaining parts of these paragraphs that concern the potentialities of human thought versus the potentialities of computing machines, already signaled above in their discussion of "calculating machines":

> It may very well be the case that the human brain is itself a "machine" with built-in limitations of its own, and that there are mathematical

[14] That was in the article, "Über eine bisher noch nicht benützte Erweiterung des finiten Standpunktes," *Dialectica*, XII (1958): 280–87; reprinted with facing English translation in Gödel, *Collected Works, Volume II: Publications 1938–1974*, pp. 240–51. As a sign of his concern with the issues involved, Gödel worked on a revision of that until late in his life, 1972, "On an Extension of Finitary Mathematics Which Has Not Yet Been Used," also in Gödel, *Collected Works, Volume II*, pp. 271–80. For the full story, see my forthcoming piece, "Lieber Herr Bernays! Lieber Herr Gödel! Gödel on Finitism, Constructivity and Hilbert's Program," in *Horizons of Truth* (Gödel centenary conference, Vienna, April 27–29, 2006).

> problems which it is incapable of solving. Even so, the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines.[15]

And (from the third paragraph)

> The discovery that there are formally indemonstrable arithmetic truths does not mean that there are truths which are forever incapable of becoming known, or that a mystic intuition must replace cogent proof. It does mean that the resources of the human intellect have not been, and cannot be, fully formalized, and that new principles of demonstration forever await invention and discovery .... Nor do the inherent limitations of calculating machines constitute a basis for valid inferences concerning the impossibility of physico-chemical explanations of living matter and human reason. The possibility of such explanations is neither precluded nor affirmed by Gödel's incompleteness theorem. The theorem does indicate that in structure and power the human brain is far more complex and subtle than any nonliving machine yet envisaged (*ibid.*).

And finally,

> Gödel's work is a remarkable example of such complexity and subtlety. It is an occasion not for dejection because of the limitations of formal deduction but for a renewed appreciation of the powers of human reason (*ibid.*).

Here again Gödel would have reacted with a "been there, done that" annoyance, since he had already laid out his thoughts in this direction fifteen years earlier in what is usually referred to as his Gibbs lecture, "Some Basic Theorems on the Foundations of Mathematics and Their Implications."[16] But once more this is something he had never published, though he wrote of his intention to do so soon after delivering the lecture; in fact it never appeared in his lifetime. After Gödel died, the text languished with a number of other important essays and lectures in his *Nachlass* until it was retrieved by our editorial group for publication in Volume III of the Gödel *Collected Works*.

[15] Nagel and Newman, "Gödel's Proof," p. 1695.

[16] Gödel's lecture was the twenty-fifth in a distinguished series set up by the American Mathematical Society to honor the nineteenth 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. See Gödel, *Collected Works, Volume III: Unpublished Essays and Lectures*, Feferman et al., eds. (New York: Oxford, 1995), pp. 304–23.

There are essentially two parts to the Gibbs lecture, both drawing conclusions from the incompleteness theorems. The first part concerns the potentialities of mind versus machines for the discovery of mathematical truths, and it is that part that we should compare with Nagel and Newman's reflections. 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. George Boolos wrote a very useful introductory note to both parts of the Gibbs lecture in Volume III of the Gödel *Works* (*ibid.*, pp. 290–304); more recently I have published an extensive critical analysis of the first part, under the title "Are There Absolutely Unsolvable Problems? Gödel's Dichotomy,"[17] and I shall be drawing on that in the following.

What I call Gödel's dichotomy is the following statement that he highlighted in the first part of the Gibbs lecture:

> *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, italics Gödel's).

By a *diophantine problem* is meant a proposition of elementary number theory (a.k.a. first order arithmetic) of a relatively simple arithmetical form whose truth or falsity is to be determined; its exact description is not important to us. 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.[18] According to Gödel, his dichotomy is a "mathematically established fact" which is a consequence of the incompleteness theorem. All that he says by way of an argument for it is the following, however:

> [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;

[17] *Philosophia Mathematica*, Series III, XIV (2006): 134–52.

[18] In modern terms, consistency statements belong to the class $\prod_1$, that is, are of the form $\forall x R(x)$ with R primitive recursive.

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. By *objective mathematics* Gödel means the totality of true statements of mathematics, which includes the totality of true statements of first-order arithmetic. The assertion that objective mathematics is incompletable is at first sight simply a consequence of the second incompleteness theorem in the form that for any consistent formal system S containing a certain basic system S_0 of (true) arithmetic, the number-theoretical statement $\mathrm{Con}(S)$ that expresses the consistency of S is true but not provable in S.

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 $\mathrm{Con}(S)$. 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 the basic system of arithmetic S_0 has been humanly accepted). Also, Gödel's sketch to establish his dichotomy should be modified slightly, as follows: either humanly demonstrable mathematics is *contained* in some consistent formal system S, or not. If it is, then $\mathrm{Con}(S)$ is an absolutely undecidable (diophantine) statement. If not, then for each consistent formal system S, there is a humanly provable statement that is not provable in S, that is, the "human mind … infinitely surpasses the powers of any finite machine."

Note well that Gödel's dichotomy is not a strict one as it stands; Gödel himself asserts that "the case that both terms of the disjunction are true is not excluded, so that there are, strictly speaking, three alternatives." This could happen, for example, if the human mind infinitely surpasses finite machines with respect to certain diophantine problems, but not with respect to *all* of them. And if we drop the word 'diophantine' from its statement, it might be that the human mind could settle all arithmetical problems, but not all problems of higher

mathematics. In fact, some logicians conjecture that Cantor's Continuum Problem is an absolutely undecidable problem of that type.

Be that as it may, how does Gödel's conclusion differ from that of Nagel and Newman? They speak of calculating machines "as currently constructed and planned, [that] operate in obedience to a fixed set of directives built in, and [that] involve mechanisms which proceed in a step-by-step manner," where Gödel speaks more precisely of Turing machines. Still, to be charitable, I think that is a reasonable interpretation of what Nagel and Newman had in mind. Second, they do not make the connection between formal systems and calculating machines, where Gödel sees these as amounting to the same thing.[19] But the essential difference is that they seem to come down in favor of the first, anti-mechanist, disjunct of the dichotomy when they say that "[by] Gödel's incompleteness theorem, there is an endless set of problems in elementary number theory for which such machines are inherently incapable of supplying answers, however complex their built-in mechanisms may be and however rapid their operations" and "the human brain appears to embody a structure of rules of operation which is far more powerful than the structure of currently conceived artificial machines"; furthermore, in the third paragraph they say that "the resources of the human intellect have not been, and cannot be, fully formalized, and ... new principles of demonstration forever await invention and discovery."

There is a lot of evidence outside of the Gibbs lecture that Gödel was also convinced of the anti-mechanist position as expressed in the first disjunct of his dichotomy. That is 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*,[20] and then at greater length in *A Logical Journey: From Gödel to Philosophy*.[21] 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, he there says:

> [It is possible that] the human mind (in the realm of pure mathematics) *is* equivalent to a finite machine that, however, is unable to understand completely its own functioning (*op. cit.*, p. 309).

[19] Calculating machines are assimilated more closely to axiomatic systems in the concluding reflections of Nagel and Newman, *Gödel's Proof*, p. 100.

[20] Wang, *From Mathematics to Philiosophy* (New York: Routledge and Kegan Paul, 1974), pp. 324–26.

[21] Wang, *A Logical Journey: From Gödel to Philosophy* (Cambridge: MIT, 1996), especially chapter 6.

And in a related footnote, despite his views concerning the "impossibility of physico-chemical explanations of ... human reason" he says that

> [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.*).

And in the next footnote, he says:

> [T]he physical working of the thinking mechanism could very well be completely understandable; the insight, however, that this particular mechanism must always lead to correct (or only consistent) results would surpass the powers of human reason (*ibid.*, p. 310).

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].*[22]

III. CRITIQUING THE MINDS VERSUS MACHINES DEBATE

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. The first outright such claim was made by J.R. Lucas in his 1961 article, "Minds, Machines and Gödel."[23] Both Benacerraf and Putnam soon objected to his argument on the basis that Lucas was assuming it is known that one's mind is consistent. Lucas, in response, has tried to shift the burden to the mechanist: "The consistency of the machine is established not by the mathematical ability of the mind,

[22] Kreisel, "Which Number-theoretic Problems Can Be Solved in Recursive Progressions on $\prod^1_1$ Paths through O?" *Journal of Symbolic Logic*, XXXVII (1972): 311–34, see p. 322; italics added.

[23] Lucas, "Minds, Machines, and Gödel," *Philosophy*, XXXVI (1961): 112–37.

but on the word of the mechanist," a burden that the mechanist can refuse to shoulder by simply citing his empirical defense.[24]

Roger Penrose is the other noted defender of the Gödelian basis for anti-mechanism, most notably in his two books, *The Emperor's New Mind*,[25] and *Shadows of the Mind*.[26] 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 mathematical truth" (*ibid.*, p. 76). But later in that work after a somewhat involved discussion, he came up with a new argument purported to show that the human mathematician cannot even consistently *believe* that his mathematical thought is circumscribed by a mechanical algorithm (*ibid.*, sections 3.16 and 3.23). Extensive critiques have been made of Penrose's original and new arguments in an issue of the journal *PSYCHE*, to which he responded in the same issue.[27] And more recently, Stewart Shapiro[28] and Per Lindström[29] have carefully analyzed and then undermined 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 cannot be done in mathematics by specific procedures, for example, 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 human mind 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,"[30]

[24] Lucas, "Minds, Machines, and Gödel: A Retrospect," in P.J.R. Millican and A. Clark, eds., *Machines and Thought: The Legacy of Alan Turing*, Volume 1 (New York: Oxford, 1996), pp. 103–24.

[25] Penrose, *The Emperor's New Mind* (New York: Oxford, 1989).

[26] Penrose, *Shadows of the Mind* (New York: Oxford, 1994).

[27] Penrose, "Beyond the Doubting of a Shadow," *Psyche*, II, 1 (1996): 89–129; also at http://psyche.cs.monash.edu.au/v2/psyche-2-23-penrose.html.

[28] Shapiro, "Mechanism, Truth, and Penrose's New Argument," *Journal of Philosophical Logic*, XXXII (2003): 19–42.

[29] Lindström, "Penrose's New Argument," *Journal of Philosophical Logic*, XXX (2001): 241–50, and "Remarks on Penrose's 'New Argument'," *Journal of Philosophical Logic*, XXXV (2006): 231–37.

[30] Shapiro, "Incompleteness, Mechanism, and Optimism," *Bulletin of Symbolic Logic*, IV (1998): 273–302.

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,[31] 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 diophantine statements of the form Con(S) for S an arbitrary effectively presented formal system need be considered. 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 com-

[31] Cf. Penrose, "Beyond the Doubting of a Shadow," pp. 137ff.

putations. But the point of that idealization was to give the strongest *negative* results, to show that certain kinds of problems cannot 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?

Here is where I think something new has to be said, something that I already drew attention to in my article on Gödel's dichotomy,[32] but that needs to be amplified. 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, 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 nondeterministic Turing machine*, and *not* with the set of its enumerable statements (even though that can equally well be supplied by a deterministic Turing machine).

One could no more disprove this modified version of the idealized mechanist's thesis than the version considered by Gödel and the others, 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. And the experience of the mathematical practitioner certainly supports the conclusion drawn by Nagel and Newman that "mathematical proof does not coincide

[32] Feferman, "Are There Absolutely Unsolvable Problems? Gödel's Dichotomy."

with the exploitation of a formalized axiomatic method," even if that cannot be demonstrated unassailably as a consequence of Gödel's incompleteness theorems.

IV. ONE WAY TO STRADDLE THE MECHANIST AND ANTI-MECHANIST POSITIONS

As I see it, the 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 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 $(=, 0, ', +, .)$, and that of ZF is fixed by taking its basic symbols to be those for equality and membership $(=, \in)$. This forces axiom schemata that may be used in such systems, such as induction in arithmetic and separation in set theory, to be infinite 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*, that is, 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.[33]

The idea is familiar from standard presentations of propositional and predicate logic, where we have such axioms as

$$P \wedge Q \rightarrow P \quad \text{and} \quad (\forall x)P(x) \rightarrow P(a),$$

[33] Cf. Feferman, "Gödel's Program for New Axioms: Why, Where, How and What?" in *Gödel '96*, P. Hájek, ed., *Lecture Notes in Logic*, VI (1996): 3–22, and Feferman, "Open-ended Schematic Axiom Systems (abstract)," *Bulletin of Symbolic Logic*, XII (2006): 145, and Feferman and Thomas Strahm, "The Unfolding of Non-finitist Arithmetic," *Annals of Pure and Applied Logic*, CIV (2000): 75–96.

and rules of inference such as

$$P, P \to Q \vdash Q \quad \text{and} \quad P \to Q(x) \vdash P \to (\forall x)\, Q(x) \text{ (for } x \text{ not free in } P).$$

We do not conceive of logic as applying to a single subject matter fixed once and for all, but rather to any subject in which we take it that we are dealing with well-defined propositions and predicates; logic is there applied by substitution for the proposition and predicate letters in its axioms and rules of inference. Similarly to allow systems like those for arithmetic and set theory to be applicable no matter what subject matter we happen to deal with, we formulate their basic principles in schematic form such as the *Induction Axiom Scheme*

$$P(0) \wedge (\forall x)[P(x) \to P(x')] \to (\forall x) P(x)$$

in the case of arithmetic, and the *Separation Scheme*

$$(\forall a)(\exists b)(\forall x)[x \in b \leftrightarrow x \in a \wedge P(x)]$$

in the case of set theory. (The idea for the latter really goes back to Zermelo's conception of the *Aussonderungsaxiom* as applicable to any "definite" predicate.)

But why, it may be (and is) asked, do I insist on the vague idea of an open-ended language for mathematics? Aren't all mathematical concepts defined in the language of set theory? It is indeed the case that the current concepts of working ("pure") mathematicians are with few exceptions 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.[34] Other outliers are to be found on the constructive fringe of mathematics in the schools of Brouwerian intuitionism and Bishop's constructivism[35] 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

[34] Cf. Saunders Mac Lane, *Categories for the Working Mathematician* (Berlin: Springer, 1971), and Feferman, "Categorical Foundations and Foundations of Category Theory," in R.E. Butts and J. Hintikka, eds., *Logic, Foundations of Mathematics and Computability Theory*, Volume I (Dordrecht: Reidel, 1977), pp. 149–65, and Feferman, "Enriched Stratified Systems for the Foundations of Category Theory," in G. Sica, ed., *What Is Category Theory?* (Monza, Italy: Polimetrica, 2006), pp. 185–203.

[35] Cf. Michael Beeson, *Foundations of Constructive Mathematics* (Berlin: Springer, 1985).

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.[36] 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. Sustained mathematical reasoning had its origins in ancient Greece some 2500 years ago, and mathematical concepts of number and space have undergone considerable evolution since then. Yet even the earliest results have permanent validity, though not necessarily as originally conceived. While Euclidean geometry is no longer considered to be the geometry of actual space, its consequences as an axiomatic development (suitably refined through the work of Pasch, Hilbert, and others), such as the angle sum theorem and Pythagoras's theorem, are as valid now within that context as in Euclid's time. On the other hand, the origins of number theory as represented in Euclid's *Elements*, including the existence of infinitely prime numbers and the fundamental theorem of arithmetic, retain their direct interpretation and validity. But the language of mathematical practice (often mixed with physical concepts) then and through the many following centuries up to the present obviously cannot be identified with the language of set theory. One thing that would instead account for the continuity of mathematical thought throughout its history is the employment of certain underlying formal patterns, such as those indicated above, that could be instantiated by the evolving concepts that have come to fill out mathematics in the process of its development. And there is no reason to believe that this evolutionary process has come to an end; it would be foolish to believe that only that which can be expressed, say, in the language of set theory, will count as mathematics henceforth.

On this picture, in order to straddle the mechanist/anti-mechanist divide at the level considered here, one will have to identify *finitely many basic forms of mathematical reasoning* which work in tandem to fully

[36] Feferman, "Does Mathematics Need New Axioms?" *Bulletin of Symbolic Logic*, VI (2000): 401–13.

constrain and distinguish it. These would constitute the mechanist side of the picture, while the openness as to what counts as a mathematical concept would constitute the anti-mechanist side. The evidence for such would have to be empirical, by showing how typical yet substantial portions of the mathematical corpus are accounted for in those terms while giving special attention to challenging cases. I have suggested steps in that direction,[37] but the program is ambitious and I have only made a start; spelling that out is planned for a future publication. In the meantime, the program itself should be treated as highly speculative, yet—I hope—worthy of serious consideration.

Considered more broadly and apart from the tendentious terms of the mechanism/anti-mechanism debate, I think the goal should be to give an informative, systematic account at a theoretical level of how the mathematical mind works that squares with experience. Characterizing the logical structure of mathematics—what constitutes a proof—is just one aspect of that, important as that may be. Other aspects—the ones that are crucial in making the difficult choices that are necessary for the mathematician to obtain proofs of difficult theorems—such as the role of heuristics, analogies, metaphors, physical and geometric intuition, visualization, and so on, have also been taken up and are being pursued in a more or less systematic way by mathematicians, philosophers, and cognitive scientists.[38] Without abandoning a basic naturalist stance, I see no usable reductive account of the mathematical experience anywhere on the horizon at the neurophysiological level let alone more basic physico-chemical levels of the sort contemplated by Nagel and Newman and currently sought by Penrose, among others.

SOLOMON FEFERMAN

Stanford University

[37] As indicated in "Are There Absolutely Unsolvable Problems?"

[38] Cf., for example, Efraim Fischbein, *Intuition in Science and Mathematics* (Dordrecht: Reidel, 1987); George Lakoff and Rafael E. Núñez, *Where Mathematics Comes From* (New York: Basic, 2000); Paolo Mancosu, Klaus Frovin Jørgensen, Stig Andur Pedersen, eds., *Visualization, Explanation and Reasoning Styles in Mathematics* (Dordrecht: Springer, 2005); and George Pólya, *Mathematics and Plausible Reasoning*, Volumes 1 and 2 (Princeton: University Press, 1968, 2nd edition) among others. In addition, there is a massive amount of anecdotal evidence for the nonmechanical essentially creative nature of mathematical research; for a small sample with further references, cf., for example, Philip J. Davis, and Reuben Hersh, *The Mathematical Experience* (Boston: Birkhäuser, 1981), and David Ruelle, *The Mathematician's Brain* (Princeton: University Press, 2007), and Benjamin H. Yandell, *The Honors Class: Hilbert's Problems and Their Solvers* (Natick, MA: A.K. Peters, 2002).