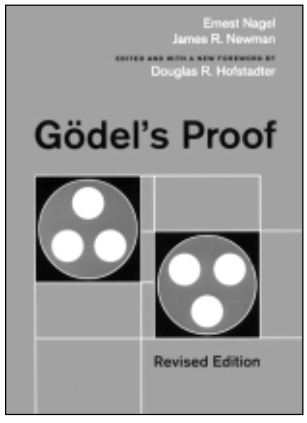
Gödel's Proof

Reviewed by Timothy McCarthy

Gödel's Proof
Ernest Nagel and James R. Newman
Revised Edition, edited and with a new foreword by Douglas R. Hofstadter
New York University Press, 2001
125 pages, $17.95
ISBN 0-1847-5816-9

In *A Mathematician's Apology*, G. H. Hardy wrote that no book is wholly bad that fires the imagination of clever young people. When *Gödel's Proof* first appeared forty-four years ago, it was the first introductory treatment of Gödel's incompleteness theorems in print and as such had an enormous formative influence on a whole generation of students. One of these was Douglas Hofstadter, the editor of the present revised edition of the book and author of the kaleidoscopic *Gödel, Escher, Bach: An Eternal Golden Braid*; another, a few years later, was the author of this review. This welcome republication of Nagel and Newman's book, with a helpful new preface and a number of substantive corrections, provides a chance to look back at a modern classic in light of the subsequent development of its subject.

The first six chapters of the book canvass the background to Gödel's paper of 1931. This paper establishes, for each formal axiomatic system of a certain not-very-special sort, the existence of sentences which are *undecidable* in the system (sentences in the language of the system which are neither provable nor refutable in the system). The paper also outlines an argument showing the unprovability in each such system of a certain sentence that may be interpreted as expressing its consistency. After a brief introductory chapter, in Chapter 2 the authors sketch the history of the consistency problem as it arose in the nineteenth century. The distinction between relative and absolute consistency proofs is introduced: a relative consistency proof is essentially a structure-preserving interpretation of one theory in another; an absolute consistency proof for a theory is an outright demonstration that no contradiction is derivable in the theory. An interpretation of Riemannian in Euclidean geometry is given, providing a historically important example of a relative consistency proof. The erroneous claim in the original edition that a model of Riemannian geometry results by taking *points* to be points on the surface of a Euclidean sphere and *lines* to be geodesics is corrected by construing the Riemannian points to be pairs of antipodal points (thus ensuring that two points determine a line). The chapter concludes with Russell's paradox, concerning the set of all non-self-membered sets,



Timothy McCarthy is professor of philosophy at the University of Illinois, Urbana-Champaign. His email address is tgmccart@uiuc.edu.

which dramatizes the consistency problem for naïve set theory.

Chapter 3 explores Hilbert's program for obtaining consistency proofs, which sought to provide absolute consistency proofs for fragments of classical mathematics in *finitistic* terms. For Hilbert "finitistic" meant, very roughly, *strictly constructive*. A finitistic consistency proof for a theory T containing arithmetic will specify an effective procedure for showing, for any proof σ in T, that σ is *not* a proof of the sentence $0 = 1$. To make this into a well-defined mathematical problem, T is to be construed as a purely formal object, characterized by a finite formal lexicon and by effective procedures (algorithms) for determining whether a finite sequence of symbols from the lexicon constitutes a formula and whether a finite sequence of formulas constitutes a proof. The systematic codification of formal logic underlying such a construal is described in Chapter 4, and Chapter 5 presents an example of a successful absolute consistency proof which is "finitistic" in the relevant sense: the authors show that a formulation of sentential logic is consistent by showing that every provable formula is a tautology (essentially by induction on the length of proofs). In general, the exposition in the first five chapters of the book is clear, concise, and correct, and especially helpful to the beginner.

In Chapter 6 Nagel and Newman begin to lay the groundwork for their exposition of Gödel's 1931 paper (the bulk of the exposition occurs in Chapter 7). The authors rightly point to a similarity between Gödel's proof and certain semantical paradoxes, pursuing a somewhat strained analogy between Gödel's construction and the Richard Paradox.[1] The idea behind Gödel's proof is also aptly compared to the Liar Paradox. A *Liar sentence* is a sentence of the form "p is not true", where "p" designates the sentence itself; such a sentence thus expresses its own falsity. Gödel showed, somewhat analogously, how to produce a sentence that expresses its own *unprovability*. This chapter goes on to usefully explore ways in which one structure may encode information about another under a mapping. It is this sort of mapping—whereby talk about the arithmetical properties of integers covertly becomes talk about the syntactic properties of a formal theory—that crucially underlies Gödel's construction of the sentence referred to above. It was in Chapter 7, which contains the exposition of the fine structure of Gödel's argument, where the original edition of the book ran into the most trouble.[2] Some of these shortcomings have been corrected in the present edition, but some substantial expository problems remain. In what follows I shall touch on what I take to be the most important of these.

Nagel and Newman clearly describe the process now called "arithmetization of syntax", whereby the syntactic objects of a formal system are coded by positive integers (Gödel numbers) and syntactic relations between them are represented by number-theoretic relations of a sort Gödel called "recursive" [*rekursiv*], now called "primitive recursive". Gödel's original target was a theory he called "PM", a somewhat simplified version of *Principia Mathematica*. PM is essentially a simple type theory, built over elementary arithmetic; that is to say, the theory introduces disjoint types of objects, sets of objects, sets of sets of objects, etc., beginning with the natural numbers as the initial type. The authors state a result they call "The Correspondence Lemma" (p. 73): that each "primitive recursive truth" can be expressed by a sentence of a certain form that is provable in PM and that each provable sentence of this form is true. There is a slight misstatement here: the term "primitive recursive" does not properly apply to propositions or sentences, but rather to certain relations (or functions). It would have considerably clarified the subsequent exposition to simply state the basic result as Gödel does: that each, say, binary primitive recursive relation R can be associated with an open sentence $\psi(x, y)$ in the language of PM such that if a pair $< k_1, k_2 >$ of integers is related by R, then the assertion $\psi(k_1, k_2)$ is provable in PM, and otherwise this assertion is refutable in PM. (A parallel result may be given for relations of any arity.) This is the property that is now called "numeral-wise expressibility"; it is a readily understandable rendering of the authors' assertion that primitive recursive notions can be "mirrored" inside PM.

We can now arrive at the heart of the argument. Gödel showed how to define a primitive recursive binary relation R such that for any natural numbers n and m, nRm iff (i) there is an open sentence ϕ (containing just the variable y, say) such that m codes the sentence resulting by substitution of the Gödel number of ϕ itself for y in ϕ, and (ii) n is *not* the Gödel number of a proof in PM of the sentence with code m. Let $\mathbf{R}(x, y)$ be the formula of PM that expresses the relation R in the above sense, and let $B(y)$ be the formula $(\forall x)\mathbf{R}(x, y)$. Finally, let k be the Gödel number of $B(y)$. The sentence G that Gödel showed to be undecidable in PM is just the assertion $B(k)$, which, as the authors observe,

[1] *The analogy was suggested by Gödel himself. First presented by Jules Richard in 1905, this paradox considers an injective enumeration $C_1, C_2, \ldots$ of conditions expressible in English. Now define a condition, R, that applies to exactly the conditions C_n such that C_n does not apply to n. Let R be C_p. Then obviously R applies to* **p** *if and only if it does not.*

[2] *For a discussion of the difficulties in the original edition, see the review by Hilary Putnam in* Philosophy of Science **29** *(April 1960).*

expresses a truth of arithmetic if and only if that sentence itself is not a theorem of PM.[3] At this point, however, matters become somewhat muddled. First, the authors attribute to Gödel a proof of the equivalence "G is provable in PM iff $\neg G$ is provable in PM" (p. 99), but then immediately add a footnote asserting, correctly, that this is not what Gödel actually proved. The equivalence is rather attributed to Rosser, who first established the incompleteness of axiomatic theories extending arithmetic on the hypothesis of simple consistency (the condition that if a sentence A is derivable in the theory, then $\neg A$ is not). The authors say that they are using the equivalence attributed to Rosser in order to simplify their exposition; the difficulty is that Rosser did not in fact prove this equivalence! Rosser's undecidable sentence is not the statement G but another statement based on a somewhat more complicated diagonal construction. Rosser's argument is, if anything, *less* simple than Gödel's, not more so; the point of Rosser's theorem is that it gets along with a *weaker hypothesis* than the one assumed by Gödel's.

What is frustrating about this expository situation is that, given the admirable job of stage-setting the authors have done, a sketch of Gödel's actual argument would have been very easy to give. Gödel did in fact establish that the sentence G is not derivable in PM on the assumption that PM is simply consistent, and the authors correctly outline this part of the argument. The argument that $\neg G$ is not derivable in PM is based on the somewhat stronger hypothesis that PM satisfies a condition Gödel called "ω-consistency", saying that for no open sentence $A(x)$ is each sentence of the form $A(n)$ provable in PM while the sentence $(\forall x)A(x)$ is refutable in PM. This is equivalent to the condition that if PM proves each individual natural number falls under a property, then it cannot prove that there are natural numbers that do not. Gödel's argument is quite simple. The hypothesis of ω-consistency ensures that of simple consistency, and so the first half of the argument shows that G is not derivable in PM. So no number codes a proof of it; whence for each n we have nRk. Thus, since the formula $\mathbf{R}(x, y)$ expresses the relation R, for each n the assertion that $\mathbf{R}(n, k)$ holds is provable in PM. By ω-consistency, then, the sentence $\neg G$ cannot be a theorem of PM, since G is the assertion that $(\forall x)\mathbf{R}(x, k)$.

In the concluding part of Chapter 7, Nagel and Newman turn to Gödel's second incompleteness theorem, concerning the unprovability of consistency, and its significance for Hilbert's program. With the exception of some misstatements coming toward the end, this material is clear and helpful. The essential observation here is that the argument of the first half of the first incompleteness theorem, that the sentence G is not provable in PM if in fact PM is consistent, can easily be transcribed into PM by means of a suitable arithmetization of the notion of proof. Gödel represents the consistency claim by the sentence $(\exists x)\neg P(x)$, where $P(x)$ is a certain formula of arithmetic which is true of precisely the codes of the sentences provable in PM. This sentence is thus an arithmetization of the assertion that there is some unprovable sentence, which is easily seen to be equivalent to simple consistency. Let 'CON' abbreviate this arithmetical consistency statement. By mimicking the argument of Gödel's first theorem, we can derive the sentence "If CON, then G holds" *in PM*. Thus, since G is not derivable in PM (assuming PM consistent), neither is CON.

Nagel and Newman subscribe to the standard interpretation of the significance of this result for Hilbert's program but are rather more careful about expressing that interpretation than a number of subsequent writers. Nagel and Newman say that the result excludes the possibility of a finitistic consistency proof for PM *that can be mirrored in PM*, but add that "no one today appears to have a clear idea of what a finitistic proof would be like that is *not* capable of being mirrored inside *Principia Mathematica*" (p. 109, n. 39). Let us say that a formal theory is *finitistically complete* if every finitistically provable arithmetical statement is derivable in the theory. So the authors' claim is that Gödel's second theorem shows that no finitistic consistency proof is available for PM *if in fact PM is finitistically complete*; they then provisionally adopt the hypothesis that PM *is* finitistically complete, and detach the conclusion that no finitistic consistency proof for PM is available. They are surely on strong ground in making this completeness claim for PM; on the most influential analysis of finitism, the finitistically provable number-theoretic statements are identified with the theorems of primitive recursive arithmetic, a very weak fragment of PM. The difficulty lies rather with the first claim: Gödel's second theorem does not, unfortunately, by itself show quite what the authors take it to show. The difficulty is that Gödel shows explicitly only that *one* arithmetic transcription of the consistency claim for PM is underivable in PM (assuming PM consistent). But even if PM is finitistically complete, the question arises whether the syntactic assertion that PM is consistent, which Gödel showed to be unprovable in PM under one arithmetical transcription of its content, may yet be provable under another.

The authors do not discuss this question, but it is interesting to observe that their stance toward

[3] *Here and in what follows, if $P(x)$ is a condition and n an integer, by $P(n)$ we mean the sentence resulting from $P(x)$ by replacing x by a standard numeral for n; the sentence thus says that n satisfies $P(x)$.*

Gödel's second theorem makes possible a short answer to it, and one favorable to their desires. The question is whether there could be an alternative arithmetic transcription of the consistency assertion for PM that *is* derivable in PM. Suppose we had such an alternative candidate, say χ. Although it is not clear in general what χ must do in order to constitute an acceptable representation of the consistency claim, from the standpoint of finitistic consistency proofs, we clearly want to require χ to be a finitistically meaningful number-theoretic statement that can be recognized to be equivalent to the unarithmetized consistency claim by finitistic means: let us say in this case that the two statements are *finitistically equivalent*. Gödel's canonical arithmetization of the consistency claim for PM certainly satisfies this condition. Both Gödel's consistency sentence and the deviant sentence χ, then, are finitistically equivalent to the unarithmetized syntactic claim of consistency for PM. They are thus finitistically equivalent as number-theoretic statements. But the authors' assumption is that PM is finitistically complete. Thus this equivalence must already be demonstrable in PM. It follows that χ is not derivable in PM; else, Gödel's consistency statement would also be derivable in PM. In short, on the authors' hypothesis about PM, *if one finitistically acceptable arithmetic transcription of the consistency claim for PM is unprovable in PM, then they are all unprovable in PM.*

Nagel and Newman recognize that the impossibility of a finitistic consistency proof for a theory does not exclude the possibility of a more general sort of meta-mathematical consistency proof for it. Thus, for example, Gentzen's consistency proof for number theory (1936) is constructive in a reasonably strong sense, exploiting, in addition to finististic devices, transfinite induction over Cantor's ordinal ϵ_0. The authors cite Gentzen's work, but rather oddly construe its significance:

> Meta-mathematical arguments establishing the consistency of formal systems such as PM have, in fact, been devised, notably by Gerhard Gentzen, a member of the Hilbert school, in 1936, and by others since then. These proofs are of great logical significance, because they propose new forms of meta-mathematical constructions, and thereby help make clear how the class of rules of inference has to be enlarged if the consistency of PM and related systems is to be established. (pp. 107–108)

But Gentzen's proof did not target PM—nor, indeed, anything approximating PM—but only elementary number theory, and although Gentzen-style consistency proofs have subsequently been provided for stronger systems, using induction on larger and larger proof-theoretic ordinals, these proofs are conducted in standard set theory and are easily transcribed into systems such as PM. They cannot, therefore, *apply* to systems such as PM.

This misunderstanding of Gentzen-type consistency proofs appears to be connected to some of the reflections with which the book concludes. In the boldest of these, the authors stand at the beginning of a series of writers who have claimed that Gödel's incompleteness theorem bears negatively on the thesis of mechanism in the philosophy of mind. Gödel's results, the authors say, show 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" (p. 112). The end-product of formalization is a *formal system*, but a formal system is essentially an algorithm for generating its theorems. We could thus express this negative claim as the thesis that *the theorem-proving capacity of the human mind cannot be represented by a Turing machine.* The argument for this surprising contention is presumably something like the following: suppose that T is a Turing machine that can prove just the number-theoretic statements I can prove. Then via Gödel's procedure I can construct a true number-theoretic proposition (on the model of the sentence G above) that T cannot prove *but that I can recognize to be true.* But then T cannot, after all, prove all the number-theoretic statements I can prove!

In his review of the original edition of the book, Hilary Putnam provided the following rather decisive response to this argument:

> Given an arbitrary machine T, all I can do is find a proposition U such that *I* can prove
>
> (*) If T is consistent, then U is true,
>
> where U is undecidable in T if in fact T is consistent. However, T can perfectly well prove (*) too! And the statement U, which T *cannot* prove (assuming consistency), *I* cannot prove either (unless I can prove that T is consistent, which is unlikely if T is very complicated)!

It follows that if there is a Turing machine T that "represents" me in the present sense, then I cannot recognize that T is *sound* (i.e., proves only true sentences). But of course there is no reason to suppose that the soundness of my entire battery of theorem-proving strategies should be transparent to *me*! I suspect that Nagel and Newman may have been led to an insouciantly optimistic attitude toward the justification of consistency claims by

their interpretation of Gentzen's result discussed above; it may have been their view that suitable extensions of Gentzen's procedure would facilitate consistency proofs for (consistent) theories of arbitrary strength.

However it arises, the authors' idea that there are nondemonstrative sources of mathematical justification has antecedents in Gödel's work and has surfaced repeatedly in the subsequent literature. But I want to emphasize that this idea by itself is entirely compatible with mechanism. At one point Gödel speculated that "it is conceivable that every proposition expressible in set theory is decidable from the present axioms plus some true assertion about the largeness of the universe of sets," and in his Gibbs Lecture he considered a strategy for introducing new axioms of the latter sort.[4] This strategy is *ampliative* (in the philosophical sense of *not deductive*; in particular, the combination of Zermelo-Fraenkel set theory with such an axiom implies the consistency statement for each previous system of this sort), but Gödel believed that the axioms in question are implicit in the general concept of set. However, such a strategy may be ampliative and yet be *effective* if in fact the choice of new axiom is effectively determined by the theory to which it is added.

Ampliative inferences are now modeled in what are called "nonmonotonic reasoning systems". In such a system there is a dynamic notion of *projection* for sentences: a sentence projected in one state of the system need not be projected in later states. The notion of *proof* is replaced by an appropriate stability property, which in the simplest of these models consists in a sentence's being projected in all descendents of a given state. When such a model is applied to number-theoretic problems, one can ask whether an analogue of Gödel's incompleteness theorem holds. Rather surprisingly, the answer is that it depends on the precise form of the theorem considered. The theorem of Rosser mentioned above turns out *not* to generalize to this context without restriction: there are nonmonotonic systems which are *effective* (in the sense that the evolution of the states of the system is described by a Turing machine), which satisfy analogues of both the consistency *and completeness* conditions, and in which each theorem of Peano Arithmetic is "provable" (in the relevant sense). But a form of Gödel's theorem itself, using a nonmonotonic analogue of Gödel's ω-consistency condition, holds for a wide variety of such systems. Stronger results are obtainable by reference to nonmonotonic analogues of what are called *reflection principles*, a condition requiring that a sentence of a certain form is provable only if it is true.

The significance of these (relatively recent) developments for Nagel and Newman's position is somewhat mixed. On the one hand, the nonmonotonic models formally implement a coherent conception of nondeductive justification in mathematics and show how it may be that the number-theoretic truths that can be justified in this way need not be effectively enumerable (the range of a computable function). On the other hand, in doing so they exhibit a computational architecture for such a conception and so undermine the idea that there is any tension between these observations and a "computationalist" understanding of mechanism. Nagel and Newman were only the first to allege a connection between Gödel's theorems and the thesis of mechanism.[5] The idea that there is such a connection was circulating at the time the original edition of the book appeared, perhaps inspired by some occasional remarks of Gödel himself, and it has come up repeatedly in the subsequent literature. But no convincing philosophical argument establishing such a connection has ever been given.

Despite its shortcomings, Nagel and Newman's book should be recognized as a classic piece of expository literature. It can still be recommended as an excellent introduction to the background of Gödel's incompleteness theorem and to the philosophical issues to which Gödel's result is connected. Some of its technical exposition is flawed; some of its philosophical claims are suspect. But it is notoriously difficult to say something philosophically sensible in this area, and there are now a number of excellent introductory presentations of the incompleteness theorems, written for people with only a modicum of logical background, containing reasonably complete proofs. Nagel and Newman's book can most profitably be read preparatory to, or in conjunction with, one of these. Its greatest merits are that it conveys the intellectual landscape in which Gödel's paper appeared, the general nature of the ideas involved in that paper, and something of the significance of the conclusions he reached. The book is an excellent point of departure. After reading it, one will want to learn more.

[4] *"Remarks before the Princeton bicentennial conference on problems in mathematics" and "Some basic theorems on the foundations of mathematics and their implications" (the Gibbs Lecture),* Collected Works*, vol. 3, Oxford University Press, 1995.*

[5] *In his foreword to the book, Douglas Hofstadter also finds the supposed connection to be unsupportable for reasons broadly similar to those given here. He has not, however, made any changes in the philosophical parts of Nagel and Newman's text.*