There can never be surprises in logic. —Ludwig Wittgenstein

Do I contradict myself?
Very well then I contradict myself.
(I am large, I contain multitudes.) —Walt Whitman

Thus, be it understood, to demonstrate a theorem, it is neither necessary nor even advantageous to know what it means. The geometer might be replaced by the "logic piano" imagined by Stanley Jevons; or, if you choose, a machine might be imagined where the assumptions were put in at one end, while the theorems came out at the other, like the legendary Chicago machine where the pigs go in alive and come out transformed into hams and sausages. No more than these machines need the mathematician know what he does.
—Henri Poincaré

4 Goedel's Proof

By ERNEST NAGEL and
JAMES R. NEWMAN

IN 1931 there appeared in a German scientific periodical an exceptionally difficult and brilliant paper entitled *"Ueber formal unentscheidbare Saetze der Principia Mathematica und verwandter Systeme"* ("On Formally Undecidable Propositions of Principia Mathematica and Related Systems"). The author of the paper was Kurt Goedel, then a young mathematician of 25 at the University of Vienna, now a member of the Institute for Advanced Study at Princeton. When at a convocation in 1952 Harvard University awarded Goedel an honorary degree, the citation described his achievement as the most important advance in mathematical logic in a quarter century.

"On Formally Undecidable Propositions of Principia Mathematica and Related Systems" is a milestone in the history of modern logic and mathematics, yet probably neither its title nor its contents were at the time of its appearance intelligible to the great majority of professional mathematicians. This is not surprising. The term "undecidable propositions" may for the moment be briefly identified as the name of propositions which can be neither proved nor disproved within a given system; the *Principia Mathematica*, to which the paper referred, is the monumental three-volume treatise by Alfred North Whitehead and Bertrand Russell on mathematical logic and the foundations of mathematics. Now familiarity with the thesis and the techniques of the *Principia*, let alone with some of the questions it raised, was not in 1931 (and is not now) a prerequisite to successful research in most branches of mathematics. There were, to be sure, a number of mathematicians, chiefly under the influence of the outstanding German mathematician David Hilbert, who

were profoundly interested in these matters; but the group was small. Logico-mathematical problems have never attracted a wide audience even among those who are partial to abstract reasoning. On the other hand, to those who were able to read Goedel's paper with understanding, its conclusions came as an astounding and a melancholy revelation. For the central theorems which it demonstrated challenged deeply rooted preconceptions concerning mathematical method, and put an end to one great hope that motivated decades of research on the foundations of mathematics. Goedel showed that the axiomatic method, which mathematicians had been exploiting with increasing power and rigor since the days of Euclid, possesses certain inherent limitations when it is applied to sufficiently complex systems—indeed, when it is applied even to relatively simple systems such as the familiar arithmetic of cardinal numbers. He also proved, in effect, that it is impossible to demonstrate the internal consistency (non-contradictoriness) of such systems, except by employing principles of inference so complex that their internal consistency is as open to doubt as that of the systems themselves. Goedel's paper was not, however, exclusively negative in import. It introduced a novel technique of analysis into the foundations of mathematics that is comparable in fertility with the power of the algebraic method which Descartes introduced into the study of geometry. It suggested and initiated new problems and branches of logico-mathematical research. It provoked a critical reappraisal, not yet completed, of widely held philosophies of knowledge in general, and of philosophies of mathematics in particular.

Despite the novelty of the techniques Goedel introduced, and the complexity of the details in his demonstrations, the major conclusions of his epoch-making paper can be made intelligible to readers with even limited mathematical preparation. The aim of the present article is to make the substance of Goedel's findings generally understandable. This aim will perhaps be most easily achieved if the reader is first briefly reminded of certain relevant developments in the history of mathematics and modern formal logic.

I

The nineteenth century witnessed a tremendous expansion and intensification of mathematical research. Many fundamental problems that had long withstood the best efforts of earlier thinkers received definitive solutions; new areas of mathematical study were created; and the foundations for various branches of the discipline were either newly laid, or were recast with the help of more rigorous techniques of analysis. In particular, the development of the non-Euclidean geometries stimulated the revision and completion of the axiomatic basis for many mathematical systems;

and axiomatic foundations were supplied for fields of inquiry which hitherto had been cultivated in a more or less intuitive manner. One important conclusion that emerged from this critical examination of the foundations of mathematics was that the traditional conception of mathematics as the "science of quantity" is both inadequate and misleading. For it became evident that mathematics is the discipline *par excellence* which draws necessary conclusions from any given set of axioms (or postulates), and that the validity of the inferences drawn does not depend upon any particular interpretation which may be assigned to the postulates. Mathematics was thus recognized to be much more "abstract" and "formal" than had been traditionally supposed. The postulates of any branch of demonstrative mathematics are not inherently "about" space, quantity, or anything else; and any special meaning which may be associated with the "descriptive" terms (or predicates) in the postulates plays no essential role in the process of deriving theorems. The sole question which confronts the pure mathematician (as distinct from the scientist who employs mathematics in investigating a special subject matter) is not whether the postulates he assumes or the conclusions he deduces from them are *true*, but only whether the alleged conclusions are in fact the *necessary logical consequences* of the initial assumptions. For example, among the undefined terms employed by Hilbert in his famous axiomatization of geometry are the following: "point," "line," "plane," "lies on," and "between." The customary meanings attributed to these (predicate) expressions undoubtedly promote the cause of discovery and learning. That is, because of the very familiarity of these notions, they not only motivate and facilitate the formulation of axioms, but they also suggest the goals of inquiry, i.e., the statements one wishes to establish as theorems. Nevertheless, as Hilbert states explicitly, for mathematical purposes familiar connotations are to be banished and the "meanings" of the expressions are to be taken as completely described by the axioms into which they enter. In more technical language the expressions are "implicitly defined" by the axioms and whatever is not embraced by the implicit definitions is irrelevant to the demonstration of theorems. The procedure recalls Russell's famous epigram: pure mathematics is the subject in which we do not know what we are talking about, nor whether what we are saying is true.

This land of rigorous abstraction, empty of all familiar landmarks, was certainly not easy to get around in. But it offered compensations in the form of a new freedom of movement and fresh vistas. The intensified formalization of mathematics emancipated men's minds from the restrictions which the standard interpretation of expressions placed on the construction of novel systems of postulates. As the meaning of certain terms became more general, less explicit, their use became broader, the inferences to be drawn from them less confined. Formalization led in fact to

a great variety of axiomatized deductive systems of considerable mathematical interest and value. Some of these systems, it must be admitted, did not lend themselves to an intuitively obvious interpretation, but this fact caused no alarm. Intuition, for one thing, is an elastic faculty; our children will have no difficulty in accepting as intuitively obvious the paradoxes of relativity, just as we do not boggle at ideas which were regarded as wholly unintuitive a couple of generations ago. Moreover, intuition, as we all know, is not a dependable guide: it cannot be used safely as a criterion of either truth or fruitfulness in scientific explorations.

A more serious problem, however, was raised by the increased abstractness of mathematics. This turned on the question whether a given set of postulates underlying a new system was internally consistent, so that no mutually contradictory theorems could be deduced from the set. The problem does not seem pressing when a set of axioms is taken to be "about" a definite and familiar domain of objects; for then it is not only significant to ask, but it may be possible to ascertain, whether the axioms are indeed true of these objects. Thus, since the Euclidean axioms were generally supposed to be true statements about space (or objects in space), apparently no mathematician prior to the nineteenth century ever entertained the question whether a pair of contradictory theorems might not some day be deduced from the axioms. The basis for this confidence in the consistency of Euclidean geometry was the sound principle that logically incompatible statements cannot be simultaneously true; accordingly, if a set of statements are true (and this was generally assumed to be the case for the Euclidean axioms), they are also mutually consistent.

But the non-Euclidean geometrics were clearly in a different case. For since their axioms were initially regarded as being plainly false of space, and, for that matter, doubtfully true of anything, the problem of establishing the internal consistency of non-Euclidean systems was recognized to be both substantial and serious. In Riemannian geometry, for example, the famous parallel postulate of Euclid (which is equivalent to the assumption that through a given point in a plane just one parallel can be drawn to a given line in the plane) is replaced by the assumption that through a given point in a plane *no* parallel can be drawn to a given line in the plane. Now suppose the question: is the Riemannian set of postulates consistent? They are evidently not true of the ordinary space of our experience. How then is their consistency to be tested? How can one prove they will not lead to contradictory theorems?

A general method was devised for solving this problem. The underlying idea is to find a "model" (or interpretation) for the postulates so that each postulate is converted into a true statement about the model. In the case of Euclidean geometry, as we have seen, the model was ordinary space. Now the method was extended to find other models, the elements of which could be used as crutches for the abstractions of the postulates. The procedure goes something like this. Suppose the following set of postulates is given concerning two classes K and L, whose special nature is left undetermined except as "implicitly" defined in the postulates:

(1) Any two members of K are contained in just one member of L.

(2) No member of K is contained in more than two members of L.

(3) The members of K are not all contained in a single member of L.

(4) Any two members of L contain just one member of K.

(5) No member of L contains more than two members of K.

From this little set, using customary rules of inference, theorems can be derived. For example, it can be shown that K contains just three members. But is the set a consistent one, so that mutually contradictory theorems can never be derived from it? The fact that no one has as yet deduced such theorems does not settle the question, because this does not prove that contradictory theorems may not eventually be deduced. The question is readily resolved, however, with the help of the following model. Let K be the class of points constituting the vertices of a triangle, and L the class of lines constituting its sides; and let us understand the phrase "a member of K is contained in a member of L" to mean that a point which is a vertex lies on a line which is a side. Each of the five abstract postulates is then converted into a true statement—for example, the first asserts that any two points which are vertices of the triangle lie on just one line which is a side. Thereby the set is proved to be consistent.

In a similar fashion the consistency of plane Riemannian geometry can be established. Let us interpret the expression "plane" in the Riemannian postulates to signify the surface of a Euclidean sphere, the expession "point" to signify a point on this surface, the expression "straight line" to signify an arc of a great circle on this surface, and so on. Each Riemannian postulate is then converted into a truth of Euclid. For example, on this interpretation the Riemannian parallel postulate reads as follows: Through a point on the surface of a sphere, no arc of a great circle can be drawn parallel to a given arc of a great circle.

All this is very tidy, no doubt, but we must not become complacent. For as any sharp eye will have seen by now we are not so much answering the problem as removing it to familiar ground. We seek to settle the question of Riemannian consistency by appealing, in effect, to the authority of Euclid. But what about his system of geometry—are its axioms consistent? To say that they are "self-evidently true," and therefore consistent, is today no longer regarded as an acceptable reply. To describe the axioms as inductive generalizations from experience would be to claim for them only some degree of probable truth. A great mass of evidence might be adduced to support them, yet a single contrary item would destroy their title of universality. Induction therefore will not suffice to establish the consistency of Euclid's geometry as logically certain. A different approach was tried by David Hilbert. He undertook to interpret the Euclidean postulates in a manner made familiar by Cartesian co-ordinate geometry, so that they are transformed into algebraic truths. Thus, in the axioms for plane geometry, construe the expression "point" to signify a pair of real numbers, the expression "straight line" to signify the relation between real numbers which is expressed by a first degree equation with two unknowns, the expression "circle" to signify the relation between numbers expressed by a quadratic equation of a certain form, and so on. The geometric statement that two distinct points uniquely determine a straight line is then transformed into the algebraic truth that two pairs of real numbers uniquely determine a linear form; the geometric theorem that a straight line intersects a circle in at most two points, is transformed into the algebraic theorem that a linear form and a quadratic form of a certain type determine at most two pairs of real numbers; and so on. In brief, the consistency of the Euclidean postulates is established by showing that they are satisfied by an algebraic model.

This method for establishing consistency is powerful and effective. Yet it too remains vulnerable to the objections set forth above. In other words the problem has again been solved in one domain only by transferring it to another. Hilbert's proof of the consistency of his postulates simply shows that *if* algebra is consistent, then so is his geometric system. The proof is merely *relative* to the assumed consistency of some other system and is not an "absolute" proof.

In attempting to solve the problem of consistency one notices a recurrent source of difficulty. It is encountered whenever a non-finite model is invoked for purposes of interpretation. It is evident that in making generalizations about space only a very limited portion—that which is accessible to our senses—serves as the basis of grand inferences; we extrapolate from the small to the universal. But where the model has a finite number of elements the difficulty is minimized, if it does not completely vanish. The vertex-triangle model used above to show the consistency of the five abstract K and L class postulates is finite; it was therefore comparatively simple to determine by actual inspection whether all the elements in the model actually satisfied the postulates. If this condition is fulfilled they are "true" and hence consistent. To illustrate: by examining in turn all the vertices of the model triangle one can learn whether any two of them lie on one side—so that the first postulate is established as true. Unfortunately, however, most of the postulate systems that constitute the foundations of important branches of mathematics cannot be mirrored in finite models and can be satisfied only by non-finite ones. One of the postulates, for example, in a well known axiomatization of elementary arithmetic asserts that *every* integer has an immediate successor which differs from any integer preceding it in the progression. It is evident that the set of postulates containing this one cannot be interpreted by means of a finite model; the model itself will have to mirror the infinity of elements postulated by the axioms. The truth (and so the consistency) of the set cannot therefore be established by inspection and enumeration. Apparently then we have reached an impasse. Finite models suffice to establish the consistency of certain sets of postulates, but these are of lesser importance. Non-finite models, necessary for the interpretation of most postulate systems, can be described only in general terms, and we are not warranted in concluding as a matter of course that the descriptions themselves are free from a concealed contradiction.

It may be tempting to suggest at this point that we can be assured of the consistency of descriptions which postulate non-finite models, if the basic notions employed in such descriptions are transparently "clear" and "certain." But the history of thought has not dealt kindly with the doctrine of intuitive knowledge which is implicit in the suggestion. In certain areas of mathematical research, in which assumptions about infinite domains play central roles, radical contradictions (or "antinomies") have turned up, despite the "intuitive" clarity of the notions involved in the assumptions, and despite the seemingly consistent character of the intellectual constructions performed. Such antinomies have emerged in the theory of transfinite numbers developed by Georg Cantor in the nineteenth century; and the occurrence of these contradictions has made plain that the apparent clarity of even such an elementary notion as that of *class,* does not guarantee the consistency of the system built on it. Now the theory of classes (or aggregates) is often made the foundation for other branches of mathematics, and in particular for elementary arithmetic. It is therefore pertinent to ask whether antinomies similar to those encountered in the theory of transfinite numbers may not infect other parts of mathematics.

In point of fact, Russell constructed a contradiction within the framework of elementary logic itself, a contradiction which is the precise analogue of the antinomy first developed in the Cantorian theory of transfinite numbers. Russell's antinomy can be stated as follows: Classes may be divided in two groups: those which do not, and those which do contain themselves as members. A class will be called "normal" if, and only if, it does not contain itself as a member. Otherwise it is "non-normal." An example of a normal class is the class of mathematicians, for patently the class itself is not a mathematician and is therefore not a member of itself. An example of a non-normal class is the class of all thinkable things; for the class of all thinkable things is itself a thinkable thing and is therefore a member of itself. Now let "N" by definition stand for the class of *all* normal classes. We ask whether N itself is a normal class. If N is normal, it is a member of itself for, by definition of "N," N is to include all normal classes; but in that case also N is non-normal because by definition of "non-normal," non-normal classes are those which contain themselves as members. On the other hand, if N is non-normal, then again it is a member of itself by definition of "non-normal," but then also it is normal because it belongs to N which is defined as normal. N, in other words, is normal if and only if N is non-normal. It follows that the statement "N is normal" is both true and false. This fatal contradiction results from an uncritical use of the apparently pellucid notion of class.

Moreover, additional antinomies were found subsequently, each of them constructed by means of familiar and seemingly cogent modes of reasoning. But the intellectual construction and formulation of non-finite models generally involves the use of possibly inconsistent sets of postulates. Accordingly, although the classical method for establishing the consistency of axioms continues to be an invaluable mathematical tool, that method does not supply a final answer to the problem it was designed to resolve.

II

The inadequacies of the model method of demonstrating consistency, and the growing apprehension, based on the discovery of the antinomies, that established mathematical systems were infected by contradictions, led to new attacks upon the problem. An alternative to relative proofs of consistency was proposed by Hilbert. He sought to construct so-called "absolute" proofs of freedom from contradiction. These we must explain briefly as a further preparation for discussing Goedel's proof.

The first requirement of an absolute proof as Hilbert conceived it is the *complete formalization* of the system. This, the reader will recall, means draining the expressions occurring within the system of any meaning whatever; they are to be regarded simply as empty, formal signs. How these signs are to be manipulated is then to be set forth explicitly in a set of rules. The purpose of this procedure is to construct a calculus which conceals nothing, which has in it only that which we intended to put in it. When theorems of this calculus are derived from the postulates by the combination and transformation of its meaningless signs in accord with precisely stated rules of operation, the danger is eliminated of the use of any unavowed principles of reasoning. Formalization is a difficult and tricky business, but it serves a valuable purpose. It reveals structure and function in naked clarity as does a cut-away working model of a machine. When a system has been formalized the logical relations between mathematical propositions are exposed to view; one is able to see the structures of configurations of certain "strings" (or sequences) of "meaningless" signs, how they hang together, are syntactically combined, nest in one another and so on.

A page covered with the "meaningless" marks of such a formalized mathematics does not *assert* anything—it is simply an abstract design or mosaic possessing a determinate structure. But suppose we as observers wish to make statements *about* a given configuration in the calculus, for example, that one "string" is longer than another or that one "string" is made up of three others. Such statements are evidently meaningful, and are expressed in a language belonging not to the calculus (or to mathematics) but to what Hilbert called "meta-mathematics" (or the language about mathematics). Meta-mathematical statements are statements *about* the signs in a calculus. They describe the kinds and arrangements of such signs when they are combined to form longer strings of marks called "formulas," and the relations between formulas in consequence of the rules of manipulation that have been specified for them. The following table illustrates some of the differences between expressions within arithmetic (mathematics) and statements about such expressions (meta-mathematics).

Mathematics	*Meta-mathematics*
For every x, if x is a prime and $x > 2$, then x is odd.	'x' is a numerical variable. '2' is a numerical constant. 'prime' is a predicate expression. '$>$' is a binary predicate.
$2 + 3 = 5$	If the sign '$=$' occurs in an expression which is a formula of arithmetic, the sign must be flanked on both its left and right sides by numerical expressions. '$2 + 3 = 5$' is a formula.
$x = x$ $0 = 0$	The formula '$0 = 0$' is derivable from the formula '$x = x$' by substituting the numeral '0' for the numerical variable 'x'.
$0 \neq 0$	'$0 \neq 0$' is not a theorem. Arithmetic is consistent—that is, it is not possible to derive from the axioms of arithmetic both the formula '$0 = 0$' and the formula '$0 \neq 0$'.

It is worth observing that, despite appearances to the contrary, the meta-mathematical statements in the right-hand column do not actually contain any of the mathematical expressions listed in the left-hand column. The right-hand column contains only the *names* of some of the arithmetical expressions in the left-hand column. This is so, because the rules of English grammar require that no English sentence shall contain the *objects* to which it refers, but only their *names*. The rule is enforced in the above table through the convention of enclosing an expression within single quotation marks in order to obtain a name for that expression. In consonance with this convention, it is correct to say that $2 + 3$ is identical with 5, but it is false to say that '$2 + 3$' is identical with '5'.

The importance of the division between the mathematical and the meta-mathematical language cannot be overemphasized. By erecting a separate, formal calculus whose symbols are free of all hidden assumptions and intuitive associations, and each of whose operations are precisely and rigidly defined, we have an instrument which exposes to plain view the nature of mathematical reasoning. But as human beings who wish to analyze this stark symbolism and to communicate our findings, we must construct another language which will enable us to describe, discuss, explain and theorize about the more formal system. Thus we separate the theory of the thing from the thing itself and devise the discourse of meta-mathematics.

It was by the application of this meta-mathematical language that Hilbert hoped to prove the consistency of the formalized calculus itself. Specifically, he sought to develop a theory of proof (*Beweistheorie*) that would yield demonstrations of consistency by an analysis of the purely structural features of expressions in uninterpreted calculi. Such an analysis consists exclusively of noting the kinds and arrangements of signs in formulas, and of showing whether a given combination of signs can be obtained from others in accordance with the explicitly stated rules of operation. An essential requirement for demonstrations of consistency, as propounded in the original version of Hilbert's program, is that they employ only *finitary* notions, and make no reference either to an infinite number of formulas or to an infinite number of operations upon them. A proof of the consistency of a set of postulates which conforms to these requirements is called "absolute." Such a proof achieves its objective by means of a bare minimum of inferential principles, without assuming the consistency of some other set of axioms. An absolute proof of the consistency of arithmetic, if one could be devised, would afford a demonstration, by finitary meta-mathematical means, that two "contradictory" formulas, such as '$(0 = 0)$' and '$\sim (0 = 0)$'—where the sign '$\sim$', called a tilde, signifies negation—are not both derivable from the axioms or initial formulas of the system, when the derivations conform to the stated rules of inference.

It may be useful, by way of illustration, to compare meta-mathematics as a theory of proof with the theory of some game, such as chess. Chess is a game played with 32 pieces of specified design on a square board containing 64 square subdivisions, where the pieces may be moved in accordance with fixed rules. The game can obviously be played without assigning any "interpretation" to the pieces or to their various positions on the board, although it is clear that such interpretations could be supplied if desired. There is thus an analogy between the game and a formalized mathematical calculus. The pieces and the squares of the board correspond to the elementary signs of the calculus; the permitted configurations of pieces on the board correspond to the formulas of the calculus; the initial positions of pieces on the board correspond to the axioms or initial formulas of the calculus; the subsequent configurations of pieces on the board correspond to formulas derived from the axioms (i.e., to the theorems); and the rules of the game correspond to the rules of derivation for the calculus. Again, although configurations of pieces on the board, like the formulas of the calculus, are "meaningless," statements about these configurations, like meta-mathematical statements about formulas, are quite meaningful. A meta-chess statement may assert, for example, that there are 20 possible opening moves for White, or that, given a certain configuration of pieces on the board with White to move, Black is mate in three moves. It is pertinent to note, moreover, that general meta-chess theorems can be established, whose proof involves the consideration of only a finite number of permissible configurations on the board. The meta-chess theorem about the number of possible opening moves for White can be established in this way; and so can the meta-chess theorem that if White has only two Knights and the King, and Black only his King, it is impossible for White to force a mate against Black. These and other meta-chess theorems can thus be proved by finitary methods of reasoning, consisting in the examination in turn of each of a finite number of configurations that can occur under stated conditions. The aim of Hilbert's theory of proof, similarly, was to demonstrate by such finitary methods the impossibility of deriving certain formulas in a calculus.

III

There are two more bridges to cross before entering upon Goedel's proof itself. Something needs be said about how and why the *Principia Mathematica* came into being; also we must give a short illustration of the formalization of a deductive system—we shall take a fragment of *Principia*—and how its consistency can be established.

Ordinarily, even when mathematical proofs conform to accepted standards of professional rigor, they suffer from one important omission. They employ principles (or rules) of inference which are not explicitly formulated, and of which mathematicians are frequently unaware. Take as example Euclid's proof that there is no greatest prime number. This is cast in the form of a *reductio ad absurdum* argument and runs as follows. Suppose there is a greatest prime x. Then:

(1) x is the greatest prime number.

(2) Form the product of all primes less than or equal to x and add 1 to the product. This yields a new number y, where $y = (2 \times 3 \times 5 \times 7 \ldots \times x) + 1$.

(3) Now if y is itself a prime, then x is not the greatest prime, for y is greater than x.

(4) But suppose y is composite, i.e., not a prime; then again x is not the greatest prime. For if y is composite, it must have a prime divisor z, which is different from each of the primes 2, 3, 5, 7 . . . x; hence z itself is a prime greater than x.

(5) But y is either prime or composite, and in either case x is not the greatest prime.

(6) Hence, since x is not the greatest prime, and x can be *any* prime number, there is no greatest prime.

We have shown the essential steps of this proof, and we could show also—though we cannot here take the time—that a number of elementary rules of inference are essential to its development, (e.g., the "Rule of Substitution," the "Rule of Detachment") and even rules and theorems belonging to more advanced parts of logical theory (e.g., the theory of "quantification," having to do with the proper use of expressions such as "all," "every," "some" and their synonyms). It has been pointed out that the use of these rules and theorems is an all but unconscious process; however, even more noteworthy is the fact that the analysis of Euclid's proof which uncovers the use of these logical props depends upon advances in the theory of logic which have occurred only within the past century. Like Molière's M. Jourdain, who spoke prose without knowing it, mathematicians have been reasoning without knowing their reasons. Modern students have had to show them the real nature of the tools of their craft.

For almost 2,000 years Aristotle's codification of valid forms of deduction was widely regarded as complete and as incapable of essential improvement. As late as 1787, the German philosopher Immanuel Kant was able to say that since Aristotle, formal logic "has not been able to advance a single step, and is to all appearances a closed and completed body of doctrine." But the fact is that the traditional logic is seriously incomplete and fails to give an account of many principles of inference employed in even quite elementary mathematical reasoning, such as the above proof of Euclid. In any event, a renaissance of logical studies in modern times began with the publication in 1847 of George Boole's *The Mathematical Analysis of Logic*. The primary concern of Boole and his immediate successors was to develop a non-numerical algebra of logic, which would provide a precise algorithm for handling more general and more varied types of deductions than were covered by traditional logical principles.

Another line of inquiry, intimately related to the work of 19th-century mathematicians on the foundations of analysis, became associated eventually with the Boolean program. This new development sought to exhibit all of pure mathematics as simply a chapter of formal logic; and it received its classical embodiment in the *Principia Mathematica* of

Whitehead and Russell in 1910. Mathematicians of the 19th century succeeded in "arithmetizing" algebra and the so-called "infinitesimal calculus," by showing that the various notions employed in mathematical analysis are definable exclusively in arithmetical terms (i.e., in terms of the integers and the arithmetical operations upon them). What Russell (and, before him, the German mathematician Gottlob Frege) sought to show was that all arithmetical notions are in turn definable in terms of purely logical ideas, and that, furthermore, the axioms of arithmetic are all deducible from a small number of basic propositions certifiable as purely logical truths. For instance, the notion of *class* belongs to general logic. Two classes are defined to be "similar," if there is a one-to-one correspondence between their members, the notion of such a correspondence being specifiable in terms of other logical ideas. A class which has no members (e.g., the class of satellites of the planet Venus) is said to be "empty." Then the cardinal number 0 can be defined as the class of all classes which are similar to an empty class. Again, a class which has a single member is said to be a "unit" class (e.g., the class of satellites of the planet Earth); and the cardinal number 1 can be defined as the class of all classes similar to a unit class. Analogous definitions can be given of the other cardinal numbers, and the various arithmetical operations can also be defined in terms of the notions of formal logic. An arithmetical statement, e.g., $1 + 1 = 2$, can then be exhibited as a condensed transcription of a statement containing only expressions belonging to general logic; and such purely logical statements can be shown to be deducible from certain logical axioms, some of which will be mentioned presently.

Principia Mathematica thus appeared to advance the final solution of the problem of consistency of mathematical systems, and of arithmetic in particular, by reducing that problem to the question of the consistency of formal logic. For if the axioms of arithmetic are simply transcriptions of theorems in logic, then the question whether these axioms are consistent is immediately transposed into the problem whether the fundamental axioms of logic are consistent.

The Frege-Russell thesis that mathematics is but a chapter of logic has not won universal acceptance from mathematicians, for various reasons of detail. Moreover, as we pointed out earlier, the antinomies of the Cantorian theory of transfinite numbers can be duplicated within logic itself, unless special measures are taken to prevent such an outcome. But are the measures adopted in *Principia Mathematica* to outflank these antinomies sufficient to exclude *all* forms of self-contradictory constructions? This cannot be asserted as a matter of course. It follows that the Frege-Russell reduction of arithmetic to logic does not provide a final answer to the consistency problem—indeed, the problem simply emerges in a more general form. On the other hand, irrespective of the validity of the Frege-Russell thesis, two features of *Principia* have proved to be of inestimable value for the further study of the problem. *Principia* supplies an inclusive system of notation, with the help of which all statements of pure mathematics can be codified in a standard manner; and *Principia* makes explicit most of the rules of formal inference (eventually these rules were made more precise and complete) which are employed in mathematical demonstrations. In short, *Principia* provides the essential instrument for investigating the entire system of formal logic as an uninterpreted calculus, whose formulas are combined and transformed in accordance with explicitly stated rules of operation.

We turn now to the formalization of a small portion of *Principia,* namely, the elementary logic of propositions. The task is to convert this fragment into a "meaningless" calculus of uninterpreted signs and to show how its freedom from contradiction can be proved.

Four steps are involved. First a complete catalogue is presented of the signs to be employed in the calculus. These are its vocabulary. Second, the "Formation Rules" are laid down. These indicate the permissible combinations of the elementary signs which are acceptable as formulas (or sentences). The rules may be said to constitute the grammar of the system. Third, the "Transformation Rules" are specified. They describe the precise structure of formulas from which some other formula is derivable. Finally, certain formulas are selected as axioms (or as "primitive formulas"). They serve as foundation for the entire system. By the expression "theorems of the system" we denote all the formulas, including the axioms, which can be derived from the axioms by successively applying the Transformation Rules. By "proof" we mean a finite sequence of legitimate formulas, each of which is either an axiom or is derivable from preceding formulas in the sequence by the Transformation Rules.

For the elementary logic of propositions (often also called the "sentential calculus") the vocabulary is extremely simple. It consists of sentential variables (which stand for sentences) and are written

'p', 'q', 'r', etc.,

of sentential connectives

'~' is short for 'not'
'v' is short for 'or'
'⊃' is short for 'if . . . then'
'.' is short for 'and'

and of parentheses, used as signs of punctuation. It is convenient to define the last two connectives in terms of the first two, so that expressions containing '⊃' or '.' can be replaced by expressions containing only 'v' and '$\sim$'. For example '$p \supset q$' is defined as being simply shorthand for the slightly longer expression '$\sim p \vee q$'.[1]

The Formation Rules are so laid down that combinations of the elementary signs which would normally be called "sentences" are designated as "formulas." Accordingly, each sentential variable will count as a formula. Moreover, if S is a formula, so is its negation $\sim (S)$; and if S_1 and S_2 are formulas, so is $(S_1) \vee (S_2)$, with similar conventions for the other connectives. Two Transformation Rules are adopted. One of them, the Rule of Substitution, says that if a sentence containing sentential variables has been accepted as logically true, any formulas may be uniformly substituted for these variables, whereupon the new sentence will also be logically true. The other rule, that of Detachment, simply says that if we have two logically true sentences of the form S_1, and $S_1 \supset S_2$, we may also accept as logically true the sentence S_2.

The axioms of the calculus (essentially those of *Principia*) are the following:

1. $(p \vee p) \supset p$
2. $p \supset (p \vee q)$
3. $(p \vee q) \supset (q \vee p)$
4. $(p \supset q) \supset [(r \vee p) \supset (r \vee q)]$

Their meaning is easily understood. The second, for instance, says that a proposition (or sentence) implies that either it or some other proposition (or sentence) is true.

Our purpose is to show that this set of axioms is not contradictory; in other words that, by using the stated Transformation Rules, it is impossible to derive from the axioms any formula S together with its negation $\sim$ S.

Now it happens that '$p \supset (\sim p \supset q)$' is a theorem in the calculus. (We shall simply accept this as a fact without exhibiting the derivation.) Suppose, then, that some formula S, as well as $\sim$ S were deducible from the axioms. (The reader will recognize the *reductio ad absurdum* approach of Euclid's proof.) By substituting S for 'p' in the theorem (as permitted by the Rule of Substitution), and applying the Rule of Detachment twice, the formula 'q' would be deducible. But this immediately has the consequence that by substituting any formula whatsoever for 'q', any formula whatsoever would be deducible from the axioms. It is thus clear that if both some formula S and its contradictory $\sim$ S were deducible from the axioms, then *any* formula would be deducible. In short, if the calculus is

not consistent, *every* formula is a theorem. And likewise, if *not* every formula is a theorem (i.e., if there is at least one formula which is not derivable from the axioms), then the calculus *is* consistent. The task, therefore, is to exhibit some formula which cannot be derived from the axioms.

The way this is done is to employ meta-mathematical reasoning upon the system to be tested. We place ourselves, so to speak, *outside* the calculus and consider how theorems are generated *within* it. The actual procedure is pretty. (1) We try to find a characteristic common to all four axioms; (2) we try to show that this characteristic is "hereditary" under the Transformation Rules—i.e., that if all the axioms have this characteristic, any formula derived from them by the rules (which is to say, any *theorem*) also has it; (3) we try to exhibit a formula that does not have this characteristic. If we succeed in this triple task, we will have an absolute proof of consistency. For if the common characteristic exists and is hereditary, so that it is transmitted to all properly derived formulas, then any array of symbols which satisfies the requirements for being a formula but nevertheless does not possess the characteristic in question cannot be a theorem. That is to say, structurally it may be a formula, yet not one which could have been derived from the axioms; or to put it yet another way, since the suspected offspring (formula) lacks an invariably inherited trait of the forebears (axioms) it cannot in fact be their descendant (theorem). Furthermore, if we can find such a formula we will have established the consistency of the calculus; because, as we noted a moment ago, if the calculus were not consistent, *every* formula could be derived from the axioms, i.e., every formula would possess the characteristic and therefore be a theorem.

Let us specify a common characteristic. The trait we have in mind is that of being a *tautology*. In common parlance *tautology* is defined as the saying of a thing twice over in different words, e.g., "John is the father of Charles and Charles is the son of John." In logic, however, a tautology is defined as a statement that excludes no logical possibilities, e.g., "Either it is raining or it is not raining." The essence of a tautology is that it is "true in all possible worlds," whence it is a *truth of logic*. Now it can be shown (though we shall not turn aside to give the demonstration) with the aid of an ingenious device known as a "truth-table," that each of the four axioms of our little set is a tautology. That is to say, if each axiom is regarded as a formula made up of simpler formulas (e.g., the compound formula or sentence $p \supset (p \vee q)$ is constituted of the simple formulas 'p' and 'q'), it must be accepted as true *irrespective* of the truth or falsity of its elementary constituents. Even the skeptical reader will have no difficulty accepting the fact, for example, that axiom 1: $(p \vee p) \supset p$ is "true in all possible worlds," if he substitutes the elementary sentence "2 is a prime

[1] That is, "if p then q" is defined as short for "either not-p or q." In view of this definition, the statement "If Galileo played the lute then Galileo was a musician" is simply a slightly more compact way of rendering what is expressed by the statement "Either Galileo did not play the lute or Galileo was a musician."

number" for the sentential variable p and derives the sentence "If 2 is a prime number or 2 is a prime number, then 2 is a prime number."

It is also possible to show that the characteristic of being a tautology is hereditary under the Transformation Rules. In sum, if the axioms are tautologous, so are all the formulas derivable from them.

Having performed these two steps, we are ready for the third. We must look for a formula which from the standpoint of its vocabulary (sentential variables) and structure (the use of the connectives) belongs to our system; yet, because it does not possess the characteristic of being a tautology, cannot be a theorem (i.e., be derivable from the axioms) and therefore cannot belong to the system. We do not have to look very hard; it is easy to exhibit such a formula. For example '$p \vee q$' fits the requirements. It purports to be a gosling but is in fact a duckling; it does not belong to the family; it is a formula but it is not a theorem. There can be no doubt that it is not a tautology. Any correct interpretation shows this at once. As an illustration, we obtain by substitution for the variables in

$$`p \vee q\text{'}$$

the sentence

"Either John is a philosopher or Charles reads *Scientific American*."

Clearly this is not a truth of logic; which is to say, it is not a sentence that is true irrespective of the truth or falsity of its elementary constituents.

We have, therefore, achieved our goal. At least one formula has been found which is not a theorem. It follows, for reasons already explained, that it is not possible to derive from the axioms of the sentential calculus both a formula and its negation. We have constructed an absolute proof of the consistency of the system.

One final point must be mentioned. It has been shown that every theorem of the sentential calculus is a tautology, a truth of logic. It is natural to ask whether, conversely, every logical truth which is expressible in the vocabulary of the calculus (i.e., every tautology) is also a theorem. The answer is yes, though the proof is too long to be shown here. Since the axioms of the calculus are sufficient for generating all logical truths expressible in the system, the axioms are said to be "complete." It is frequently of paramount interest to determine whether an axiomatized system is complete. Indeed, a powerful motive for axiomatizing various branches of mathematics has been the desire to specify a sufficient set of initial assumptions from which all the true statements in some field of analysis are deducible. Thus, when Euclid axiomatized elementary geometry, he apparently selected his axioms so as to make it possible to deduce from them all geometric truths which were already established, as well as those still to be discovered. (Euclid's inclusion of his famous parallel postulate in his list of axioms showed remarkable insight. For as was subsequently proved, this postulate is not derivable from the others of the set, so that without the parallel postulate the remaining axioms are surely incomplete.) A similar objective controlled the axiomatization of elementary arithmetic toward the close of the nineteenth century. Until recently, it was assumed as a matter of course that a complete set of axioms for any given branch of mathematics could always be specified. In particular, it seems to have been generally believed that the axioms proposed for arithmetic by nineteenth-century mathematicians were in fact complete, or at worst could be made complete by the addition of a finite number of further axioms. The discovery that this is not so, is one of Goedel's achievements.

IV

The sentential calculus is an example of a mathematical system for which the objectives of Hilbert's theory of proof are fully realized. As we have pointed out, however, this calculus codifies only a fragment of formal logic, and its vocabulary and formal apparatus do not suffice to develop within its framework even elementary arithmetic. On the other hand, Hilbert's program has been successfully carried out for more inclusive systems, which have been shown to be both consistent and complete by meta-mathematical reasoning. For example, an absolute proof of consistency has been given for a system of arithmetic which allows for the *addition*, though not for the *multiplication*, of cardinal numbers. But can a system such as *Principia*, in which the whole and not merely a fragment of arithmetic is expressible, be proved consistent in the sense of Hilbert's program? Repeated attempts at constructing such a proof were unsuccessful; and the publication of Goedel's paper in 1931 showed, finally, that all such efforts are doomed to failure.

What did Goedel establish, and how did he prove his results? His main conclusions are two-fold. In the first place, he showed that no meta-mathematical proof is possible for the formal consistency of a system comprehensive enough to contain the *whole* of arithmetic; unless, that is, the meta-mathematical proof employs rules of inference whose consistency is as doubtful as is the consistency of the Transformation Rules used in deriving theorems *within* the system. But thus one dragon is slain only to create another.

Goedel's second main conclusion is even more surprising and revolutionary in its import, for it makes evident a fundamental limitation in the power of the axiomatic method. Goedel showed that *Principia*, or any other system within which arithmetic can be developed, is *essentially incomplete*. In other words, given *any* consistent set of arithmetical axioms, there are true arithmetical statements which are not derivable from the set. This essential point deserves illustration. Mathematics abounds in statements which seem self-evident, to which no exceptions have been found, which nevertheless have thwarted all attempts at proof. A simple example is Goldbach's "theorem" which states that every even number is the sum of two primes; yet no one has succeeded in finding a proof valid for *all* even numbers. Goldbach's conjecture presents us with a statement that may be true, but may not be derivable from the axioms of arithmetic. Now it may be suggested that the axioms should be modified or augmented to take care of this and related theorems by making them derivable. But Goedel has shown that this approach promises no final cure. That is, even if the given set of axioms is augmented by the addition of any finite number of arithmetical postulates, there will always be *further* arithmetic truths which are not formally derivable from the augmented set. Such further truths may, to be sure, be established by some form of meta-mathematical reasoning *about* an arithmetical system; but this procedure does not fit the requirement that the calculus must so to speak be self-contained, that the logical truths in question must be exhibited as the formal consequences of the specified axioms *within* the system. There is, it seems, an inherent limitation in the axiomatic method as a way of systematizing the whole of arithmetic.

How did Goedel prove his conclusions? Up to a point, the structure of his demonstration is modeled, as he himself noted, on the reasoning involved in one of the logical antinomies known as the "Richard Paradox," first propounded by the French mathematician, Jules Richard, in 1905. The paradox can be stated as follows. Assume some definite language (e.g., English) in which the various purely arithmetical properties of the integers can be expressed; and consider the definitions of these properties which can be formulated in the notation of that language. Thus, the property of being a prime number may be defined by: "not divisible by any integer other than one and itself"; the property of being a perfect square may be defined by: "being equal to the product of some integer by that integer"; and so on. It is easily seen that each such definition will contain a finite number of words and therefore a finite number of letters of the alphabet. This being so, the definitions can be placed in serial order, according to the number of letters they contain (definitions with the same number of letters can be arranged alphabetically under their serial tag). To each definition there will then correspond a unique integer—for example, the definition with the smallest number of letters will correspond to the number 1, the next definition in the ordered series will correspond to 2, and so on. We come now to an odd little point. Since each definition has an integer attached to it, it may happen in certain cases that an integer possesses the very property designated by the definition to which the integer is serially attached. (This is the same sort of thing that would happen if we prefixed to each of a list of English words the descriptive tags "short" or "long," and the word "short" itself appeared in the list. "Short" itself would of course have the tag "short" attached to it.) Suppose, for instance, the defining expression "not divisible by any integer other than one and itself" happens to be correlated with the number 17; then 17 has the property designated by that expression. On the other hand, suppose that the defining expression "being equal to the product of some integer by that integer" turns out to be correlated with the number 20; then 20 does not have the property designated by that expression. We shall now say that (in the second example) the number 20 has the property of being "Richardian," while (in the first example) the number 17 does not have the property of being "Richardian." More generally, we define the property of being Richardian as follows: "not having the property designated by the defining expression with which an integer is correlated in the serially ordered set of definitions." But observe that this last expression itself defines a numerical property, so that this expression also must belong to the above series of definitions. This being so, it must correspond to some number, say *n*, giving its position in the series. The question may then be posed, reminiscent of Russell's antinomy, whether the number *n* itself is Richardian. Almost at once we can see the fatal contradiction looming. For *n* is Richardian if, and only if, it does *not* possess the property designated by the definition with which *n* is correlated; and it is easy to see that therefore *n* is Richardian if and only if *n* is not Richardian. Accordingly, the statement "*n* is Richardian" is both true and false.

The contradiction can be avoided if we notice that in constructing it we have not played the game quite fairly. Pursuant to our initial stipulations, we were invited to consider the definitions expressible in a language about numbers and their arithmetical properties. However, it was not intended to consider definitions involving reference to the *notation* used in *formulating* numerical properties. In other words, the series of definitions in the above construction was supposed to include only expressions which refer exclusively to such notions as arithmetical addition, multiplication, and the like. Accordingly, the definition of being Richardian does not belong to this series, for this definition makes reference to such *meta-mathematical* notions as the number of *letters* occurring in *expressions*. The Richard Paradox can therefore be outflanked by distinguishing carefully between statements *within* arithmetic (which make no reference to any system of notation) and statements *about* the language in which arithmetic is codified.

The reasoning in the Richard Paradox is evidently fallacious. Its construction nevertheless suggests that it might be possible to "map" (or "mirror") meta-mathematical statements *about* a sufficiently comprehensive formal system *into* the system itself. If this were possible, then meta-mathematical statements about a system would be *represented* by statements within the system. Thereby one could achieve the desirable end of getting the formal system to speak about itself—a most valuable form of self-consciousness. The idea of such mapping is a familiar one in mathematics. It is employed in coordinate geometry, which translates geometric statements into algebraic ones, so that geometric relations are mapped onto algebraic ones. The idea is manifestly used in the construction of ordinary maps, since the construction consists in projecting configurations on the surface of a sphere onto a plane, so that relations between plane figures can mirror the relations between the spherical ones. The idea also plays a role in mathematical physics when, for example, relations between properties of electric currents are represented in the language of hydrodynamics. The basic fact which underlies all these mapping procedures is that an abstract structure of relations embodied in one domain of "objects" is exhibited to hold between "objects" in some other domain. In consequence, deductive relations between statements about the first domain can be established by exploring (often more conveniently and easily) the deductive relations between statements about their counterparts. For example, complicated geometrical relations between surfaces in space are usually more readily studied by way of the algebraic formulas for such surfaces. Similarly, questions about complicated logical relations between assertions may be more readily handled *via* the arithmetical representatives of those assertions.

In any event, the exploitation of this notion of mapping is the key to the argument in Goedel's revolutionary paper. In a manner suggested by the Richard Paradox, but without falling victim to the fallacy involved in its construction, Goedel showed that meta-mathematical statements *about* a formalized arithmetical calculus can indeed be represented *within* that calculus. In fact, he found a method of representation such that neither the arithmetical formula corresponding to a certain true meta-mathematical statement about the formula, nor the arithmetical formula corresponding to the denial of the statement, is demonstrable within the calculus. Since one of these arithmetical formulas must codify an arithmetical truth, but neither is derivable from the axioms, the axioms are incomplete. As we shall see, this incompleteness is incurable. Moreover, Goedel indicated how to construct an arithmetical formula to represent the meta-mathematical statement "The calculus is consistent," and he showed that this formula is not demonstrable within the calculus. Accordingly, the consistency of arithmetic cannot be established except by using rules of inference whose consistency is at least as doubtful as is the consistency of arithmetic itself.

Goedel's paper, as we said at the outset, is difficult. Forty-six preliminary definitions together with several important lemmas must be mastered before the main results are reached. We shall take a much easier road; nevertheless we hope at least to offer glimpses of the argument.

Goedel first showed that a formalized system of arithmetic can be set up in which it is possible to associate with each elementary sign, each formula (or sequence of signs) and each proof (that is, each finite sequence of formulas) a *unique integer*. This integer, a distinctive label, is called the "Goedel number" of the sign, formula or proof.

As an illustration take the following correspondence. The top row lists part of the basic vocabulary with the help of which the whole of arithmetic can be formulated. The second row lists under each of these basic signs its corresponding Goedel number.

'~'	1	
'v'	2	
'⊃'	3	
'∃'	4	(a qualifying symbol meaning "there is" or "there are," so that "$(\exists x)$" means "There is an x" or "For every x")
'='	5	
'0'	6	
's'	7	(for representing the immediate successor of a number)
'('	8	left-hand parenthesis)
')'	9	right-hand parenthesis)
','	10	(the comma)

In addition to these basic signs we also require sentential variables ('*p*', '*q*', etc.) for which sentences may be substituted, individual variables ('*x*', '*y*', etc.) for which numerals and numerical expressions may be substituted, and predicate variables ('*P*', '*Q*', etc.) for which predicates may be substituted. These are assigned Goedel numbers in accordance with the following rules.

Associate with each sentential variable an integer greater than 10 but divisible by 3; with each individual variable, an integer greater than 10 which leaves a remainder of 1 on division by 3; and with each predicate variable, an integer greater than 10 which leaves a remainder of 2 on division by 3. This done, each elementary sign of the system is now associated with a unique number.

Consider next a formula of the system, for example, '$(\exists x)(x = sy)$' (when literally translated, it reads: 'There is an x, such that x is the immediate successor of y,' and in effect says that every number has an immediate successor). The numbers associated with its ten constituent elementary signs are, respectively, 8, 4, 13, 9, 8, 13, 5, 7, 16, 9. We now agree to associate with the formula itself the number which is the product of the first ten primes in order of magnitude, each prime being raised to a power equal to the Goedel number of the corresponding elementary sign. By this convention the formula is associated with the number $2^8 \times 3^4 \times 5^{13} \times 7^9 \times 11^8 \times 13^{13} \times 17^5 \times 19^7 \times 23^{16} \times 29^9$; let us refer to the number as m. In a similar fashion, every formula can be made to correspond to a unique number.

Consider, finally, a sequence of formulas as may occur in some proof, for example, the sequence:

$$(\exists x)(x = sy)$$
$$(\exists x)(x = s0)$$

This second formula when translated reads '0 has an immediate successor'; it is derivable from the first formula by substituting '0' for the "free" variable 'y'. (A variable is said to be "free" in a formula if it is not preceded in that formula by a quantifier containing this variable. Thus, the variable 'x' is *not* free in either of these formulas.) We have already determined the Goedel number of the first formula: it is m; and suppose that n is the Goedel number of the second formula. We now agree to associate with the sequence of two formulas the number which is the product of the first two primes in order of magnitude, each prime being raised to a power equal to the Goedel number of the corresponding formula in the sequence. The above sequence is accordingly associated with the number $k = 2^m \times 3^n$. In like manner, a number is associated with each sequence of formulas. It is easy to see that following this procedure, every expression in the system can be tagged with a unique Goedel number.

What has been done so far is to establish a method for completely arithmetizing a formal system. The method is essentially a set of directions for establishing a correspondence between certain integers and the various elements or combinations of elements of the system. Once an expression is given, it can be uniquely numbered. But more than that, once a Goedel number is given, the expression it represents can be exactly analyzed or "retrieved," since the number itself, having been arrived at as a product of prime numbers, can be factored into these primes (as we know from a classic theorem of arithmetic) in only one way. In other words, we can take the number apart like a machine, see how it was constructed and what went into it; which is to say we can dissect an expression, a proof, in the same way.

This leads to the next step. We have already spoken of "mapping." Now we can extend the process with the help of the Goedel numbers so that meta-mathematical statements can be completely mirrored within the calculus; so that meta-mathematics itself becomes completely "arithmetized." In particular, every *meta-mathematical characterization* of the structure of expressions in the system, everything we say about them, is mapped into an arithmetical function of integers; and *every meta-mathematical statement about relations between formulas* is mapped into an arithmetical relation between integers. (By an *arithmetical function* is meant an expression such as $2 + 3$, $(7 \times 5) + 8$, and so on: that is, a function of an integer in itself an integer. By an *arithmetical relation* is meant a proposition—which may be true or false—such as $5 = 3$, $7 > 4$, and so on.) The importance of this arithmetization of meta-mathematics stems from the fact that, since each of its statements can be uniquely represented in the formal system by an expression tagged with a Goedel number, relations of logical dependence between meta-mathematical statements can be explored by examining relations between integers and their factors. To take a trivial analogue: if customers in a supermarket are given tickets with numbers determining the order in which they are to be waited on when buying meat, it is a simple matter, merely by scrutinizing the numbers themselves to discover (a) how many persons have been served, (b) how many are waiting, (c) who precedes whom and by how many customers, etc.

Consider the meta-mathematical statement: 'The sequence of formulas whose Goedel number is x is a demonstration for the formula whose Goedel number is z.' This statement is represented (mirrored) by a definite formula *in* the arithmetical calculus, a formula which expresses a purely arithmetical relation between x and z. (In the above example of assigning the Goedel number k to a demonstration, we found that $k = 2^m \times 3^n$; and a little reflection shows that there is a definite though complex arithmetical relation between k, the Goedel number of the proof, and n, the Goedel number of the conclusion.) We write this arithmetical relation between x and z as the formula 'Dem (x,z)', to remind ourselves of the meta-mathematical statement to which it corresponds. Similarly, the meta-mathematical statement: "The sequence of formulas with Goedel number x is *not* a demonstration for the formula with the Goedel number z," is also represented by a definite formula in the arithmetical formalism. This formula we shall write as '$\sim$ Dem (x,z)'.

We shall need one additional bit of special notation for stating the crux of Goedel's argument. Begin with an example. The formula '$(\exists x)(x = sy)$' has the Goedel number m, and the variable 'y' has the Goedel number 16. Substitute in this formula for the variable with Goedel number 16 (i.e., for 'y') the numeral for m. We then obtain the formula '$(\exists x)(x = sm)$'. This latter formula obviously also has a Goedel number—a number which can be actually calculated, and which, in fact, is a certain complex arithmetical function of the two numbers m and 16. However, instead of calculating this Goedel number, we can give an unambiguous meta-mathematical characterization for it: it is the Goedel number of the formula which is obtained from the formula with Goedel number m, by substituting for the variable with Goedel number 16 the numeral for m. Accordingly, this meta-mathematical characterization corresponds to a definite arithmetical function of the numbers m and 16, a function which can be expressed within the arithmetical calculus. We shall write this function as 'sub $(m, 16, m)$', to remind ourselves of the meta-mathematical description which it represents. More generally, the expression 'sub $(y, 16, y)$' is the mirror-image *within* the arithmetical formalism of the meta-mathematical characterization: "the Goedel number of the formula which is obtained from the formula with Goedel number y, by substituting for the variable with Goedel number 16 the numeral for y." It should be noted that, when a definite numeral is substituted for 'y' in 'sub $(y, 16, y)$', sub $(y, 16, y)$ is a definite integer which is the Goedel number of a certain formula.

We are now equipped to follow in outline Goedel's argument. Consider the formula '$(x) \sim \text{Dem}\ (x,z)$'. This represents, in the arithmetical calculus, the meta-mathematical statement "For every x, where x is the Goedel number of a demonstration, x is not the number of a demonstration for the formula whose Goedel number is z." This formula may therefore be regarded as a formal paraphrase of the statement "The formula with Goedel number z is not demonstrable." What Goedel was able to show was that a certain special case of this formula itself is in fact not formally demonstrable. To construct this special case we start with a formula which we shall display as line (1):

(1) $(x) \sim \text{Dem}\ (x, \text{sub}\ (y, 16, y))$

It corresponds to the meta-mathematical statement that the formula with the Goedel number sub $(y, 16, y)$ is not demonstrable. Moreover, since line (1) is a formula within the arithmetical calculus, it has its own Goedel number, say n. Let us now obtain another formula from the one on line (1) by substituting the numeral for n for the variable with Goedel number 16 (i.e., for 'y'). We thus arrive at the special case we wished to construct, and display it as line (2):

(2) $(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$

Since this last formula occurs within the arithmetical calculus, it must have a Goedel number. What is its Goedel number? A little reflection shows that it is sub $(n, 16, n)$. To see this, we must recall that sub $(n, 16, n)$ is the Goedel number of the formula which is obtained from the formula with Goedel number n, by substituting for the variable with Goedel number 16 (i.e., for 'y') the numeral for n. But the formula (2) has indeed been obtained from the formula with Goedel number n (i.e., from the formula on line (1)) by substituting for the variable 'y' the numeral for n. Let us also remind ourselves, however, that the formula '$(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$' is the mirror-image *within* the arithmetical calculus of the meta-mathematical statement: "The formula whose Goedel number is sub $(n, 16, n)$ is not demonstrable." It follows that the *arithmetical formula* '$(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$' *represents the meta-mathematical statement*: "The formula '$(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$' is not demonstrable." In a sense, therefore, this arithmetical formula can be construed as saying that it itself is not demonstrable.

Goedel is now able to show, in a manner reminiscent of the Richard Paradox, but free from the fallacious reasoning involved in that puzzle, that this arithmetical formula is indeed not demonstrable. The argument from this point on is relatively simple and straightforward. He shows that *if* the formula were demonstrable, then its formal contradictory (i.e., '$\sim (x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$', which in effect says that the formula *is* in fact demonstrable) would also be demonstrable; and conversely, if the formal contradictory of the formula were demonstrable, the formula itself would also be demonstrable. But as was noted earlier, if a formula as well as its contradictory can both be derived from a set of axioms, the axioms are not consistent. Accordingly, if the axioms are consistent, neither the formula nor its contradictory is demonstrable. In short, if the axioms are consistent, the formula is "undecidable"—neither the formula nor its contradictory can be formally deduced from the axioms.

Very well. Yet there is a surprise coming. For although the formula is undecidable if the axioms are consistent, it can nevertheless be shown by meta-mathematical reasoning to be true. That is to say, the formula is a true arithmetical statement which expresses a complex but definite numerical property of integers—just as the formula '$(x) \sim (x + 3 = 2)$' (in words, "There is no positive integer which when added to 3 will equal 2") expresses another but much simpler property of integers. The reasoning that shows the truth of the undecidable formula is rather simple. In the first place, on the assumption that arithmetic is consistent, we have already established the meta-mathematical statement: "The formula '$(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$' is not demonstrable." It must be accepted, then, that this meta-mathematical statement is true. Secondly, the statement is *represented* within arithmetic by that very formula itself. Third, we recall that meta-mathematical statements have been mapped upon the arithmetical formalism in such a way that true-mathematical statements always correspond to true arithmetical formulas. (Indeed, this is the whole point of the mapping procedure—just as in analytic geometry, geometric statements are mapped onto algebra in such a way that true geometric statements always correspond to true algebraic ones.) Accordingly, the formula in question must be true. We have thus established an arithmetical truth, not by deducing it formally from the axioms of arithmetic, but by a meta-mathematical argument.

When we were discussing the sentential calculus, we explained that the axioms of that system are "complete," since all the logical truths expressible in the system are formally derivable from the axioms. More generally, we can say that the axioms of any formalized system are "complete" if every true statement expressible in the system is formally deducible from the axioms. A set of axioms is therefore "incomplete" if not every true statement expressible in the system is formally derivable from them. It follows, since we have now established as true an arithmetical formula which is not derivable from the axioms of arithmetic, the system is incomplete. Moreover, the system is *essentially incomplete*, which means that even if we added this true but undemonstrable formula to the axioms as a further axiom, the augmented system would still not suffice to yield formally all arithmetical truths: another true arithmetical formula could be constructed, such that neither the formula nor its contradictory would be demonstrable within the enlarged system. This remarkable conclusion would hold, no matter how often we enlarged the system by adding further axioms to it.

We come then to the coda of Goedel's amazing and profound intellectual symphony. It can be shown that the meta-mathematical statement just established, namely, "If arithmetic is consistent, then it is incomplete," itself corresponds to a demonstrable formula in the arithmetical system. But the antecedent clause of this formula (the one corresponding to the meta-mathematical statement "arithmetic is consistent") is not demonstrable within the system. For if it were, the consequent clause of the formula (the one corresponding to the statement "arithmetic is incomplete," and which in fact turns out to be our old friend '$(x) \sim \text{Dem}\ (x, \text{sub}\ (n, 16, n))$') would also be demonstrable. This conclusion would, however, be incompatible with the previously obtained result that the latter formula is not demonstrable. The grand final step is now before us: we must conclude that the consistency of arithmetic cannot be established by any meta-mathematical reasoning which can be represented within the formalism of arithmetic!

A meta-mathematical proof of the consistency of arithmetic is not excluded by this capital result of Goedel's analysis. In point of fact, meta-mathematical proofs of the consistency of arithmetic have been constructed, notably by Gerhard Gentzen, a member of the Hilbert school, in 1936. But such proofs are in a sense pointless if, as can be demonstrated, they employ rules of inference whose *own* internal consistency is as much open to doubt as is the formal consistency of arithmetic itself. Thus, Gentzen used the so-called "principle of transfinite mathematical induction" in his proof. But the principle in effect stipulates that a formula is derivable from an *infinite* class of premises. Its use therefore requires the employment of nonfinitistic meta-mathematical notions, and so raises once more the question which Hilbert's original program was intended to resolve.

The import of Goedel's conclusions is far-reaching, though it has not yet been fully fathomed. They 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. They also show that there is an endless number of true arithmetical statements which cannot be formally deduced from any specified set of axioms in accordance with a closed set of rules of inference. It follows, therefore, that an axiomatic approach to number theory, for example, cannot exhaust the domain of arithmetic truth, and that mathematical proof does not coincide with the exploitation of a formalized axiomatic method. Just

in what way a general notion of mathematical or logical truth is to be defined which is adequate to the fact here stated, and whether, as Goedel himself appears to believe, only a thoroughgoing Platonic realism can supply such a definition, are problems still under debate and too difficult for more than mention here.

Goedel'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 Goedel'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. 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 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. There is no immediate prospect of replacing the human mind by robots.

None of this is to be construed, however, as an invitation to despair, or as an excuse for mystery mongering. 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. We have seen that mathematical propositions which cannot be established by formal deduction from a given set of axioms, may nevertheless be established by "informal" meta-mathematical reasoning. It would be an altogether irresponsible claim to maintain that the formally indemonstrable truths Goedel established by meta-mathematical arguments are asserted in the absence of any proof or by appeals simply to an uncontrolled intuition. 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 Goedel'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. Goedel's own 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 creative reason.