1 Introduction

1.1 Inexhaustibility: the positive side of incompleteness

In 1931, Kurt Gödel presented his famous incompleteness theorem, which has since had a great and continuing impact on logic and the philosophy of mathematics, and has also like no other result in formal logic caught the interest and imagination of the general public.

Gödel presented two results in his 1931 paper, usually referred to as the first and the second incompleteness theorem. The proof of the first incompleteness theorem shows that for every consistent formal axiomatic theory in a wide class of such theories, there is at least one statement which can be formulated in the language of the theory but can neither be proved nor disproved in the theory. Such a statement is said to be *undecidable* in the theory. By a consistent formal theory is meant one in which no logical contradiction — both a statement *A* and its negation not-*A* — can be proved. The second theorem states that for a wide class of such theories *T*, if *T* is consistent, the consistency of *T* cannot be proved using only the axioms of the theory *T* itself.

In discussions of the meaning and implications of these theorems, their negative or limiting aspects are most often kept in the foreground: a formal theory of arithmetic cannot be complete, a theory cannot be proved consistent using only the resources of that theory. But the second incompleteness theorem also has a positive aspect, which was emphasized by Gödel:

> It is *this* theorem [the second incompleteness theorem] which makes the incompletability of mathematics particularly evident. For, *it makes it impossible that someone should set up a certain well-defined system of axioms and rules and consistently make the following assertion about it: All of these axioms and rules I perceive (with mathematical certitude) to be correct, and moreover I believe that they contain all of mathematics*. If somebody makes such a statement he contradicts himself. For if he perceives the axioms under consideration to be correct, he also perceives (with the same certainty) that

> they are consistent. Hence he has a mathematical insight not derivable from his axioms.[1]

This positive implication of the incompleteness theorem is that we have a way of extending any system of axioms for mathematics that we recognize as correct to a logically stronger system of axioms that we will also recognize as correct, namely by adding the statement that the old system is consistent as a new axiom. That the resulting system is logically stronger than the old system means that it proves everything that the old system proves, and more besides. Thus our mathematical knowledge would appear to be *inexhaustible* in the sense that it cannot be pinned down in any one formal axiomatic theory.

This implication, or apparent implication, of the incompleteness theorem is neither uncontroversial nor unproblematic. There is something rather strange about the impossibility of pinning down which axioms we recognize as correct. Shouldn't it be possible to somehow incorporate in those axioms the very principle that any collection of valid axioms can be extended to a stronger collection, and thereby set down in axiomatic form all of our current mathematical knowledge? And if this really can't be done, do we have some sort of ineffable or unformalizable insight that cannot be fully expressed in rules? If so, what are the limits of what can be proved using this insight, and how can those limits be described? In particular, just what is it that happens when we try to systematically and ad infinitum extend a theory by adding successive consistency statements as new axioms?

These are the questions to be pursued in this book, along with some important side issues. Although the questions are philosophical rather than mathematical, we need to understand the relevant mathematical concepts and theorems in order to arrive at any good answers to the questions, and so much of the book consists of mathematical definitions and proofs.

The book is aimed at readers who wish to understand the inexhaustibility phenomenon pointed out by Gödel, whatever their level of logical expertise. No knowledge of logic or mathematics will be presupposed except some basic school arithmetic and some slight acquaintance with sets and functions. The relevant basics of predicate logic, recursion theory, and the theory of ordinals and ordinal notations will be introduced and explained in an ad hoc fashion as the exposition proceeds. It may be that the presentation is sometimes too brief to suffice as an introduction, in which case the non-expert reader will need to consult other sources for supplementary explanations or for a more extended and systematic treatment.

[1] *Collected Works*, vol. III, p.309. Italics in the original.

In a few places a mathematical illustration may be used which assumes some further mathematical background. In particular, this is true of the presentation and application of Euler's product formula, which presupposes some calculus. Nothing essential will be lost by skipping lightly over such passages.

In the philosophy of mathematics, we need a good choice of mathematical examples to work with. This doesn't mean that the examples must be difficult or advanced mathematics. It's not the case that the philosophical problems become more difficult or sophisticated as more difficult mathematics is introduced. The mathematician G.H.Hardy went so far as to proclaim (in Hardy [1929]) that "Philosophy proper is a subject, on the one hand so hopelessly obscure, on the other so astonishingly elementary, that there knowledge hardly counts." In a philosophical discussion of the apparent phenomenon of inexhaustibility, it's clear enough that some knowledge of Gödel's theorem, the properties of sequences of Gödelian extensions of theories (here based on autonomous recursive progressions of theories as conceived by Turing and Feferman) and some other technical topics treated in the book is needed, and it's also a good idea to have some understanding of mathematical proof as it exists in mathematics. However, the precise choice of definitions and results from logic and mathematics in this book has no profound justification, and to a considerable extent only reflects the author's preferences and limitations.

The book begins in Chapter 2 with a bit of arithmetic, both because some arithmetic will be used in later chapters and in order to provide some examples for the philosophical and logical discussion to refer to. There is an emphasis on some logical aspects of elementary arithmetical reasoning that will be described more formally in later chapters. Chapter 3 introduces some further examples of mathematics in the form of some classical results about primes, and also explains the informal distinction between arithmetical, elementary, and analytic proofs of arithmetical theorems. The formal development of the logical results used or pondered in the philosophical argument of the book begins in Chapter 4. The material in chapters 4 to 11 covers some basic predicate logic, set theory, and recursion theory leading up to the proof of the incompleteness theorems in Chapter 12. The last four chapters, beginning with Chapter 12, form the core of the argument of the book concerning inexhaustibility. A reader with a good background in logic can go straight to Chapter 12 or 13 and consult the index and earlier chapters as needed. A reader with a not-so-good background in logic who finds some of the earlier chapters heavy going may also choose to skip forward to the later chapters, and judge on the basis of the argument presented in §15.3 whether it would be worthwhile to go to the trouble of studying the earlier chapters.

The remainder of this introductory chapter will be devoted to some general comments about the aims and claims of the book, and about the difficulties a reader may encounter in coming to grips with a rather odd mixture of technical results and philosophical argument of this sort.

1.2 Two Gödelian traditions

In translation, the title of Gödel's 1931 paper is "On formally undecidable propositions of *Principia Mathematica* and related systems I". In the paper, Gödel proved the incompleteness theorem for a particular formal system *P*, which was not that of Russell's and Whitehead's monumental work *Principia Mathematica*, but was indeed related to that system. At the end of §2 of the paper, he remarks:

> In the proof of Theorem IV [the first incompleteness theorem] no properties of the system P were used besides the following:
>
> 1. The class of axioms and the rules of inference (that is, the relation "immediate consequence") are recursively definable (as soon as we replace the primitive signs in some way by natural numbers);
>
> 2. Every recursive relation is definable (in the sense of theorem V) in the system P.
>
> Therefore, in every formal system that satisfies the assumptions 1 and 2 and is ω-consistent there are undecidable propositions of the form (x)F(x), where F is a recursively defined property of natural numbers, and likewise in every extension of such a system by a recursively definable ω-consistent class of axioms.

Gödel goes on to remark that these conditions are clearly satisfied by various formal theories used in logic and set theory.

This is a common pattern in logic, probably more so than in other formal sciences. A result is proved or a concept defined for some particular formalism, and it's "clear" how to extend or adapt it to various other related formalisms. The formalisms are typically axiomatic theories, or systematizations of the rules of logical reasoning, or formalisms for expressing algorithms. It is "clear" how to extend or adapt the result or concept only on the basis of experience with these different formalisms, together with an appreciation of what is central to the definition or result in question. One might think that it should be feasible to formulate the definitions or results and their proofs in sufficient generality to cover these different formalisms. To some extent this can be done, but on the whole it is probably correct to say that Gödel's theorem and many other results in logic are best understood by seeing

the proof carried through in detail for some particular formalism, rather than by attempting to achieve any great generality in the proof itself.

This feature carries over to philosophical discussions connected with Gödel's theorem. In this book, inexhaustibility will be discussed in connection with arithmetic and extensions of arithmetic, but it will be clear how the discussion applies, for example, in the case of set theory and extensions of set theory. Also, the consequences of extending a theory by adding consistency statements or reflection principles ad infinitum will be formally worked out for one particular approach, that of transfinite recursive progressions, but with the understanding that a similar treatment can be carried out for other approaches to be found in the literature. The philosophical and informal conclusions and arguments in the book are independent of the particular formal approach taken.

Gödel also initiated in his paper another tradition that presents a considerable difficulty for the beginning student. Gödel's proof of the second incompleteness theorem, which states that a formalization of the statement "*T* is consistent" is not provable in *T* itself, is based on the fact that the (first half of the) first incompleteness theorem for *T is* provable in *T* itself. Gödel only presented this fact as a plausible claim, pointing out that the proof of the first incompleteness theorem only used arithmetical reasoning of a kind formalizable in *T.* He intended to give a rigorous proof of the claim in part II of the paper, which was also to substantiate the earlier observations regarding the wide applicability of the incompleteness theorem. However, part II never appeared. One reason for this was that the argument Gödel gave for the second incompleteness theorem was in fact quite convincing to his readers; another reason was that a detailed formal proof of the claim made in connection with the second theorem was given by Hilbert and Bernays in their two-volume *Grundlagen der Mathematik* (1939). The formal details turned out to require quite a lot of rather tedious work. In most proofs of Gödel's theorem it is customary even today to skip these formal details when it comes to proving the second incompleteness theorem.

If one looks today at the logical literature, one finds that there is a great deal of handwaving going on of the same kind as that used by Gödel in his paper. That is, informal arguments are given to show that a formalization of a certain statement is formally provable in a certain theory. These informal arguments can be quite difficult to grasp for a non-expert, but they are convincing and clear enough to the experts, who have no doubt that they could be expanded to any level of detail required to establish the assertion formally. So why the handwaving? Because the full formal proofs would in most cases be very long and boring and not add anything to our understanding of the proof or the result proved. Of course this situation

is not found only in logic. In mathematical proofs in general, many steps are often skipped or sketched when it is clear to an expert reader how they could be carried out if it were necessary to do so. But in logic we come upon this kind of reasoning at a rather early stage in our studies, and applied in non-trivial cases. Also, because of such phenomena as ω-incompleteness (to be dealt with later in the book), there are often some subtleties involved in understanding why a proof can be formalized in one theory but not in another.

In the later chapters of this book, there will be quite a lot of handwaving of this traditional kind, in connection with the step from an ordinary or informal proof of a mathematical statement to the conclusion that there is a formal derivation in some particular theory of a formalization of the statement. However, the earlier chapters present and illustrate methods that should allow a determined reader to replace the handwaving with detailed arguments.

1.3 Truth and provability

Any difficulties that a reader may have with the mathematical or logical reasoning in a book such as this can be cleared up, by wrestling with proofs and definitions and by consulting other sources. The philosophical reasoning is another matter. It's common for philosophical writers to complain that they fail to recognize their own opinions or arguments in the critical remarks of others, and for readers to find numerous unclear points in and possible interpretations of those opinions and arguments. The comments on truth and provability in this introductory section are intended to help clarify the main argument of the book and forestall some possible misunderstandings.

Truth

Throughout the book, the exposition and argument will rely on the notion of *truth* (in the form of the predicate "is true") applied to mathematical (in particular arithmetical) statements. Thus for example it will be argued that the property of a theory T that is chiefly of interest in connection with inexhaustibility is that the theory is *sound*, meaning that all theorems of T are true.

A common reaction to philosophical arguments involving truth is to ask just what is meant by "true". Indeed this question is often asked even in a purely mathematical context, as when it is stated that Gödel's theorem proves, for a wide class of the-

ories *T*, that if *T* is consistent then there is a true arithmetical statement ϕ_T not provable in *T*. The question then arises, if ϕ_T isn't provable in *T*, in what sense is it true? Is it provable in some other theory, and if so which one? Or is it true in the sense that it can be "seen to be true"? Or is it perhaps true in the sense of there being a correspondence between ϕ_T and some mathematical reality that is independent of our knowledge?

Underlying such questions is a tendency to think that whenever we speak of the *truth* of mathematical statements (whether these statements represent axioms or theorems or open questions), we are no longer talking about mathematical matters, but have entered a realm of epistemological or metaphysical argument or speculation. Mathematicians and others who shy away from such argument and speculation will therefore often surround "true" with scare quotes or avoid using the word altogether, preferring to speak about provability or derivability rather than truth.

This association of the concept of truth with philosophical issues is a natural one, since the question "What is mathematical truth?" is often used as a label for a set of fuzzily defined but distinctly philosophical questions about the nature of mathematics. But there is also a formal and mathematical use of "true" in logic, and it would be a mistake to assume that references to truth automatically take us into the realm of philosophy. In particular, in this book the reply to all questions about what is meant by "true arithmetical statement", in any mathematical or non-mathematical context, is that being true is a *mathematically* defined property of statements in the formal language of arithmetic. Thus the statement "ϕ_T is true" is itself a mathematical statement. When we unravel the mathematical definition of "true arithmetical statement" we find that for any arithmetical statement ϕ, "ϕ is a true arithmetical statement" is mathematically equivalent to ϕ itself. Similarly for references to the truth or falsity of other formalized mathematical statements.

For a reader who is unfamiliar with formal logic, it will probably be unclear at this point what a mathematical definition of "true arithmetical statement" might look like. The latter part of Chapter 7 shows how to define truth mathematically, after the necessary mathematical and logical concepts have been introduced. Basically, what such definitions amount to is a mathematical spelling out of the explanation that a statement is true if and only if things are as they are said to be in the statement. This was called "the semantic conception of truth" by Tarski, who was the first to introduce such mathematical definitions of truth (see Tarski [1944]). The essential point regarding this use of "true", as understood in this book, can be appreciated without knowledge of any of the technical details: "ϕ is true", said of an arithmetical statement ϕ, is not a statement about what can be proved or known,

and nor is it a statement about any correspondence between ϕ and a mathematical reality. It is a mathematical statement, mathematically equivalent to ϕ itself.

This stipulation concerning the meaning of "true" applied to arithmetical and other mathematical statements is basic to the presentation in the book. Note that it is only a stipulation. I am not arguing that this is what *should* be meant by "true" in a discussion of inexhaustibility, and I am not expressing any theory about mathematical truth or about anything else by using "true" in this way. In particular, the various philosophical problems and misgivings commonly associated with the concept of mathematical truth do not go away when we define "true" mathematically so as to make ϕ and "ϕ is true" equivalent. The mathematical definitions of "true statement" (for various formally defined classes of mathematical statements) to be given in later chapters will not serve to allay any doubts or misgivings that a reader may have concerning the meaning of mathematical statements or the justification of basic mathematical principles, since the definitions themselves freely use ordinary mathematics. Although there will be some discussion of the interpretation and justification of mathematical axioms in later chapters, there is no attempt to systematically introduce and justify mathematical or logical concepts and principles from a philosophical or foundational point of view. Rather, the argument of the book introduces and uses whatever parts of logic and mathematics are seen as relevant — arithmetic, inductive definitions and proofs, ordinals, functions, sets, and so on. Readers who regard some or all of the logical and mathematical apparatus invoked in the discussion as standing in need of further explanation or justification (either generally, or in the context of a philosophical discussion of this kind) will need to decide for themselves how this affects the significance of the conclusions put forward in the book.

Provability

The question what "provable" means is less often asked than the corresponding question for "true", but is less easily answered. For the discussion in this book, it is essential to distinguish between what will be called formal provability and actual provability. These terms will next be explained.

Gödel's theorem and the extensions of that theorem considered in this book are about provability in axiomatic theories, that is to say about the *existence*, in the mathematical sense, of formal proofs in such theories. A *formal proof* of a statement *A* in a theory *T* is a mathematical structure, such as a tree or sequence, built up from statements in the formally defined language of the theory *T*, using axioms and rules of reasoning associated with *T*, and with a formalization of the statement

A as its designated conclusion. For emphasis, the term *formal provability* will sometimes be used when we are talking about a formalized statement being provable in some formal theory. Formal provability is always relative to some such theory.

That a formalization of a mathematical statement — say the statement "there are infinitely many twin primes" — is provable in a formal theory does not in itself imply that the statement can be *proved* in the ordinary mathematical sense, that is, that an argument establishing the statement as a mathematical theorem can be given. As an extreme instance, any statement is provable in a theory in which it is taken as an axiom, but this tells us nothing about whether or not the statement can be proved in the ordinary sense. Less trivially, even if the statement is provable in a formal theory embodying axioms and rules of reasoning recognized as correct in ordinary mathematics, it doesn't follow that the statement can be proved in any practical sense, simply because there is a limit on the length and complexity of proofs that human beings can produce or understand.

The term *actually provable* will be used for statements that can in fact be proved to the satisfaction of the mathematical community, that is, for which some actual argument or series of arguments establishing the statement as a theorem can be produced, studied, printed, and so on. Such arguments will be referred to as *actual proofs*. When we ask where the proof of some theorem first appeared, or ask if somebody has studied the proof, or if anybody has checked the whole proof, we are asking about the actual proof — the existing text, figures, reports of the results of computations, and so on, on the basis of which the theorem is accepted as such in the mathematical community.

While formal provability is a precise and indeed mathematically defined notion, (given some formal theory or class of formal theories), actual provability is neither precise, nor mathematically definable. Rather, it is a vague concept which depends on the consensus of some mathematical community and on the resources available to that community. The only thing that is clear about actual provability is that if a statement has in fact been proved, then it is actually provable. Even so, there are in some cases disagreements over whether a statement has been proved or not, either because of uncertainties regarding the correctness of steps (or computations) in the proof, or more fundamentally because of disagreement over which axioms or methods of reasoning are admissible in mathematical proofs.

There are two important connections between formal provability and actual provability. The first lies in the foundational role played by formal systems in mathematics. It is accepted in the mainstream mathematical community that if a statement has actually been proved, there exists a formal derivation of a formalization of that

statement in some standard formal system such as the system of axiomatic set theory ZFC. This serves two useful purposes. First, it provides a standard for what we may call *local rigor* in proofs: any uncertainty about whether or not a particular step or segment in a proof is valid can be resolved by making that part of the proof more explicit and detailed, if necessary to the point where there is no doubt that it can be formalized in some particular formal theory whose axioms and rules of inference are accepted as correct. (Note that this does not mean that it would be at all helpful or feasible to replace the entire proof by a formal proof.) The second useful purpose is to allow people to set aside, in a purely mathematical context, any disagreements over what axioms or methods of reasoning are admissible. For example, a published proof of a particular theorem can be recognized as having established that (a formalization of) the theorem is provable in ZFC, even while it remains an open question just how far that proof can be simplified and whether the theorem or its proof makes sense from some particular foundational point of view (such as finitism or intuitionism).

The second significant connection between formal and actual provability consists in the existence of *negative* results about formal provability with implications for actual provability. If it is known that there is no formal proof in a theory T of a statement A, and if the rules and axioms of T include formalized versions of some particular rules of mathematical reasoning, we can conclude that there can be no actual proof of A that uses only those rules. In particular, if A is known to be unprovable in a formal system such as ZFC, it can be concluded that A cannot be proved using the methods of "ordinary mathematics", since those methods are known to be formalizable in ZFC. Thus most mathematicians will cease to regard it as a mathematical problem to prove or disprove a statement A if A is shown to be undecidable in ZFC. Of course this does not mean that A has been shown to be undecidable in any absolute sense. On the contrary, it is perfectly possible that "ordinary mathematics" will in the future come to encompass axioms or rules that do make it possible to prove or disprove A. But finding such new rules and axioms is not part of ordinary mathematical activity, and it remains a (mathematically) highly significant fact that A cannot be settled by mathematical proof as now understood.

The comments above were made in terms of *known* negative results about formal provability. If we know that A is not provable in ZFC, we know that A cannot be proved by ordinary mathematical reasoning. But of course we can also say that whether or not we know that A is not provable in ZFC, if it is in fact not provable in ZFC, then it cannot in fact be proved by ordinary mathematical reasoning, even though we may never come to realize this. Thus it is sometimes suggested that the reason why some famous mathematical conjecture has so far withstood all attempts

to prove or disprove it may be that it is in fact undecidable in ZFC. Although such suggestions usually have nothing to support them, they cannot be dismissed on general grounds.

Now consider *positive* results about formal provability. If we *know* that A is provable in some theory T whose rules and axioms we accept as valid, we will not necessarily thereby be able to produce an actual proof of A that can be formalized in T (for reasons already touched on, and elaborated below). We will however be able to prove A on the basis of the knowledge that it is provable in T by invoking the principle that every statement provable in T is true. This is an application of a *reflection principle* for T to produce an actual proof of a theorem which is formally provable in T without the reflection principle. We may come to know that A is provable in T for example by using computers to search for a formal proof of A in T. The formal proof found by the computer may be too long or complex to be at all intelligible to us, but by using the reflection principle we can conclude that A is true. Such arguments, which include reports of the results of computations that go far beyond what has traditionally been used in mathematical proofs, are actual proofs in the sense defined, although to some extent controversial, since proofs that do not appeal to the results of computations that cannot be carried out by hand are generally considered more satisfactory. But reflection principles for T can also be used to prove theorems that are *not* provable in T itself — in particular, every reflection principle for T has the consequence that T is consistent — and for this reason they will play a central role in later chapters.

While the mere fact of A being unprovable in ZFC had consequences for actual provability (whether or not we are able to establish that A is unprovable in ZFC), the mere fact of a statement A being provable in ZFC or in any theory T whose rules and axioms we recognize as valid does not imply that A is actually provable. This is most simply illustrated by numerical statements of the form "p is a prime", "p has exactly two prime factors", and so on. All true statements of this form are formally provable using the ordinary rules of arithmetic, and indeed in a logically very weak formal theory, but as far as actual mathematics is concerned, there is no guarantee whatever, for a given such statement with p a large integer expressed in some standard notation, that it is possible to decide whether the statement is true or false. Specialists in factoring techniques may or may not succeed in doing so, using various mathematical ideas and results in combination with computer calculations. Once we have established the truth of a statement of this kind, by whatever means, we know that it is provable in a logically weak axiomatic theory T, but this does not mean that we could in fact have used only T to arrive at the truth of the statement.

The main significance of these distinctions in the present context is that the questions and arguments in this book concern only formal provability, not actual provability. (Thus for example there will be no discussion of the use of reflection principles to make possible or simplify actual proofs; for this topic, see Harrison [1995].) More precisely, they concern formal provability in arithmetic and certain extensions of arithmetic by axioms that we recognize as valid. Thus one may say that in speaking of provability in a discussion of this kind, we are speaking about what truths are implicit in axioms and rules that we recognize as valid, rather than about methods by which we actually prove or could prove mathematical theorems. The apparent inexhaustibility of our mathematical knowledge, as brought out by Gödel, is based on the discovery that the truth of certain mathematical statements ("the theory T is consistent") is *not* formally implicit in the axioms and rules of a theory, even though it is implicit in the theory being valid. Our accepting the theory as valid means that we are equally justified in accepting the formally undecidable statement as true, and so we are faced with the questions outlined in the introductory section.