Truth and Proof[a]

The antinomy of the liar, a basic obstacle to an adequate definition of truth in natural languages, reappears in formalized languages as a constructive argument showing not all true sentences can be proved

Alfred Tarski

The subject of this article is an old one. It has been frequently discussed in modern logical and philosophical literature, and it would not be easy to contribute anything original to the discussion. To many readers, I am afraid, none of the ideas put forward in the article will appear essentially novel; nonetheless, I hope they may find some interest in the way the material has been arranged and knitted together.

As the title indicates, I wish to discuss here two different though related notions: the notion of truth and the notion of proof. Actually the article is divided into three sections. The first section is concerned exclusively with the notion of truth, the second deals primarily with the notion of proof, and the third is a discussion of the relationship between these two notions.

The Notion of Truth

The task of explaining the meaning of the term "true" will be interpreted here in a restricted way. The notion of truth occurs in many different contexts, and there are several distinct categories of objects to which the term "true" is applied. In a psychological discussion one might speak of true emotions as well as true beliefs; in a discourse from the domain of esthetics the inner truth of an object of art might be analyzed. In this article, however, we are interested only in what might be called the logical notion of truth. More specifically, we concern ourselves exclusively with the meaning of the term "true" when this term is used to refer to sentences. Presumably this was the original use of the term "true" in human language. Sentences are treated here as linguistic objects, as certain strings of sounds or written signs. (Of course, not every such string is a sentence.) Moreover, when speaking of sentences, we shall always have in mind what are called in grammar declarative sentences, and not interrogative or imperative sentences.

Whenever one explains the meaning of any term drawn from everyday language, he should bear in mind that the goal and the logical status of such an explanation may vary from one case to another. For instance, the explanation may be intended as an account of the actual use of the term involved, and is thus subject to questioning whether the account is indeed correct. At some other time an explanation may be of a normative nature, that is, it may be offered as a suggestion that the term be used in some definite way, without claiming that the suggestion conforms to the way in which the term is actually used; such an explanation can be evaluated, for instance, from the point of view of its usefulness but not of its correctness. Some further alternatives could also be listed.

The explanation we wish to give in the present case is, to an extent, of mixed character. What will be offered can be treated in principle as a suggestion for a definite way of using the term "true", but the offering will be accompanied by the belief that it is in agreement with the prevailing usage of this term in everyday language.

Our understanding of the notion of truth seems to agree essentially with various explanations of this notion that have been given in philosophical literature. What may be the earliest explanation can be found in Aristotle's *Metaphysics:*

> **To say of what is that it is not, or of what is not that it is, is false, while to say of what is that it is, or of what is not that it is not, is true.**

Here and in the subsequent discussion the word "false" means the same as the expression "not true" and can be replaced by the latter.

The intuitive content of Aristotle's formulation appears to be rather clear. Nevertheless, the formulation leaves much to be desired from the point of view of precision and formal correctness. For one thing, it is not general enough; it refers only to sentences that "say" about something "that it is" or "that it is not"; in most cases it would hardly be possible to cast a sentence in this mold without slanting the sense of the sentence and forcing the spirit of the language. This is perhaps one of the reasons why in modern philosophy various substitutes for the Aristotelian formulation have been offered. As examples we quote the following:

> **A sentence is true if it denotes the existing state of affairs.**
>
> **The truth of a sentence consists in its conformity with (or correspondence to) the reality.**

Due to the use of technical philosophical terms these formulations have undoubtedly a very "scholarly" sound. Nonetheless, it is my feeling that the new formulations, when analyzed more closely, prove to be less clear and unequivocal than the one put forward by Aristotle.

The conception of truth that found its expression in the Aristotelian formulation (and in related formulations of more recent origin) is usually referred to as the *classical*, or *semantic conception of truth*. By semantics we mean the part of logic that, loosely speaking, discusses the relations between linguistic objects (such as sentences) and what is expressed by these objects. The semantic [[64]] character of the term "true" is clearly revealed by the explanation offered by Aristotle and by some formulations that will be given later in this article. One speaks sometimes of the correspondence theory of truth as the theory based on the classical conception.

(In modern philosophical literature some other conceptions and theories of truth are also discussed, such as the pragmatic conception and the coherence theory. These conceptions seem to be of an exclusively normative character and have little connection with the actual usage of the term "true"; none of them has been formulated so far with any

[a] [[Scientific American, June 1969, 63–70, 75–77]]

degree of clarity and precision. They will not be discussed in the present article.)

We shall attempt to obtain here a more precise explanation of the classical conception of truth, one that could supersede the Aristotelian formulation while preserving its basic intentions. To this end we shall have to resort to some techniques of contemporary logic. We shall also have to specify the language whose sentences we are concerned with; this is necessary if only for the reason that a string of sounds or signs, which is a true or a false sentence but at any rate a meaningful sentence in one language, may be a meaningless expression in another. For the time being let us assume that the language with which we are concerned is the common English language.

We begin with a simple problem. Consider a sentence in English whose meaning does not raise any doubts, say the sentence "snow is white". For brevity we denote this sentence by "*S*", so that "*S*" becomes the name of the sentence. We ask ourselves the question: What do we mean by saying that *S* is true or that it is false? The answer to this question is simple: in the spirit of Aristotelian explanation, by saying that *S* is true we mean simply that snow is white, and by saying that *S* is false we mean that snow is not white. By eliminating the symbol "*S*" we arrive at the following formulations:

(1) "snow is white" is true if and only if snow is white.
(1') "snow is white" is false if and only if snow is not white.

Thus (1) and (1') provide satisfactory explanations of the meaning of the terms "true" and "false" when these terms are referred to the sentence "snow is white". We can regard (1) and (1') as partial definitions of the terms "true" and "false", in fact, as definitions of these terms with respect to a particular sentence. Notice that (1), as well as (1'), has the form prescribed for definitions by the rules of logic, namely the form of logical equivalence. It consists of two parts, the left and the right side of the equivalence, combined by the connective "if and only if". The left side is the definiendum, the phrase whose meaning is explained by the definition; the right side is the definiens, the phrase that provides the explanation. In the present case the definiendum is the following expression:

"snow is white" is true;

the definiens has the form:

snow is white.

It might seem at first sight that (1), when regarded as a definition, exhibits an essential flaw widely discussed in traditional logic as a vicious circle. The reason is that certain words, for example "snow", occur in both the definiens and the definiendum. Actually, however, these occurrences have an entirely different character. The word "snow" is a syntactical, or organic, part of the definiens; in fact the definiens is a sentence, and the word "snow" is its subject. The definiendum is also a sentence; it expresses the fact that the definiens is a true sentence. Its subject is a name of the definiens formed by putting the definiens in quotes. (When saying something of an object, one always uses a name of this object and not the object itself, even when dealing with linguistic objects.) For several reasons an expression enclosed in quotes must be treated grammatically as a single word having no syntactical parts. Hence the word "snow", which undoubtedly occurs in the definiendum as a part, does not occur there as a syntactical part. A medieval logician would say that "snow" occurs in the definiens *in suppositione formalis* and in the definiendum *in suppositione materialis*. However, words which are not syntactical parts of the definiendum cannot create a vicious circle, and the danger of a vicious circle vanishes.

The preceding remarks touch on some questions which are rather subtle and not quite simple from the logical point of view. Instead of elaborating on them, I shall indicate another manner in which any fears of a vicious circle can be dispelled. In formulating (1) we have applied a common method of forming a name of a sentence, or of any other expression, which consists in putting the expression in quotes. The method has many virtues, but it is also the source of the difficulties discussed above. To remove these difficulties let us try another method of forming names of expressions, in fact a method that can be characterized as a letter-by-letter description of an expression. Using this method we obtain instead of (1) the following lengthy formulation:

(2) The string of three words, the first of which is the string of the letters Es, En, O and Double-U, the second is the string of letters I and Es, and the third is the string of the letters Double-U, Aitch, I, Te, and E, is a true sentence if and only if snow is white.

Formulation (2) does not differ from (1) in its meaning; (1) can simply be regarded as an abbreviated form of (2). The new formulation is certainly much less perspicuous than the old one, but it has the advantage that it creates no appearance of a vicious circle.

Partial definitions of truth analogous to (1) (or (2)) can be constructed for other sentences as well. Each of these definitions has the form:

(3) "*p*" is true if and only if *p*,

where "*p*" is to be replaced on both sides of (3) by the sentence for which the definition is constructed. Special attention should be paid, however, to those situations in which the sentence put in place of "*p*" happens to contain the word "true" as a syntactical part. The corresponding equivalence (3) cannot then be viewed as a partial definition of truth since, when treated as such, it would obviously exhibit a vicious circle. Even in this case, however, (3) is a meaningful sentence, and it is actually a true sentence from the point of view of the classical conception of truth. For illustration, imagine that in a review of a book one finds the following sentence:

> (4) **Not every sentence in this book is true.**

By applying to (4) the Aristotelian criterion, we see that the sentence (4) is true if, in fact, not every sentence in the book concerned is true, and that (4) is false otherwise; in other words, we can assert the equivalence obtained from (3) by taking (4) for "p". Of course, this equivalence states merely the conditions under which the sentence (4) is true or is not true, but by itself the equivalence does not enable us to decide which is actually the case. To verify the judgment expressed in (4) one would have to read attentively the book reviewed and ana- [[65]] lyze the truth of the sentences contained in it.

In the light of the preceding discussion we can now reformulate our main problem. We stipulate that the use of the term "true" in its reference to sentences in English then and only then conforms with the classical conception of truth if it enables us to ascertain every equivalence of the form (3) in which "p" is replaced on both sides by an arbitrary English sentence. If this condition is satisfied, we shall say simply that the use of the term "true" is adequate. Thus our main problem is: can we establish an adequate use of the term "true" for sentences in English and, if so, then by what methods? We can, of course, raise an analogous question for sentences in any other language.

The problem will be solved completely if we manage to construct a general definition of truth that will be adequate in the sense that it will carry with it as logical consequences all the equivalences of form (3). If such a definition is accepted by English-speaking people, it will obviously establish an adequate use of the term "true".

Under certain special assumptions the construction of a general definition of truth is easy. Assume, in fact, that we are interested, not in the whole common English language, but only in a fragment of it, and that we wish to define the term "true" exclusively in reference to sentences of the fragmentary language; we shall refer to this fragmentary language as the language L. Assume further that L is provided with precise syntactical rules which enable us, in each particular case, to distinguish a sentence from an expression which is not a sentence, and that the number of all sentences in the language L is finite (though possibly very large). Assume, finally, that the word "true" does not occur in L and that the meaning of all words in L is sufficiently clear, so that we have no objection to using them in defining truth. Under these assumptions proceed as follows. First, prepare a complete list of all sentences in L; suppose, for example, that there are exactly 1,000 sentences in L, and agree to use the symbols "s_1", "s_2", …, "$s_{1,000}$" as abbreviations for consecutive sentences on the list. Next, for each of the sentences "s_1", "s_2", …, "$s_{1,000}$" construct a partial definition of truth by substituting successively these sentences for "p" on both sides of the schema (3). Finally, form the logical conjunction of all these partial definitions; in other words, combine them in one statement by putting the connective "and" between any two consecutive partial definitions. The only thing that remains to be done is to give the resulting conjunction a different, but logically equivalent, form, so as to satisfy formal requirements imposed on definitions by rules of logic:

> (5) **For every sentence x (in the language L), x is true if and only if either**
> s_1, **and x is identical to "s_1",**
> **or**
> s_2, **and x is identical to "s_2",**
> …
> …
> **or finally,**
> $s_{1,000}$, **and x is identical to "$s_{1,000}$".**

We have thus arrived at a statement which can indeed be accepted as the desired general definition of truth: it is formally correct and is adequate in the sense that it implies all the equivalences of the form (3) in which "p" has been replaced by any sentence of the language L. We notice in passing that (5) is a sentence in English but obviously not in the language L; since (5) contains all sentences in L as proper parts, it cannot coincide with any of them. Further discussion will throw more light on this point.

For obvious reasons the procedure just outlined cannot be followed if we are interested in the whole of the English language and not merely in a fragment of it. When trying to prepare a complete list of English sentences, we meet from the start the difficulty that the rules of English grammar do not determine precisely the form of expressions (strings of words) which should be regarded as sentences: a particular expression, say an exclamation, may function as a sentence in some given context, whereas an expression of the same form will not function so in some other context. Furthermore, the set of all sentences in English is, potentially at least, infinite. Although it is certainly true that only a finite number of sentences have been formulated in speech and writing by human beings up to the present moment, probably nobody would agree that the list of all these sentences comprehends all sentences in English. On the contrary, it seems likely that on seeing such a list each of us could easily produce an English sentence which is not on the list. Finally, the fact that the word "true" does occur in English prevents by itself an application of the procedure previously described.

From these remarks it does not follow that the desired definition of truth for arbitrary sentences in English cannot be obtained in some other way, possibly by using a different idea. There is, however, a more serious and fundamental reason that seems to preclude this possibility. More than that, the mere supposition that an adequate use of the term "true" (in its reference to arbitrary sentences in English) has been secured by any method whatsoever appears to lead to a contradiction. The simplest argument that provides such a contradiction is known as the *antinomy of the liar;* it will be carried through in the next few lines.

Consider the following sentence:[b]

> **(6) The sentence printed in red on page 65 of the June 1969 issue of *Scientific American* is false.**

Let us agree to use "s" as an abbreviation for this sentence. Looking at the date of this magazine, and the number of this page, we easily check that "s" is just the only sentence printed in red on page 65 of the June 1969 issue of *Scientific American*. Hence it follows, in particular, that

> **(7) "s" is false if and only if the sentence printed in red on page 65 of the June 1969 issue of *Scientific American* is false.**

On the other hand, "s" is undoubtedly a sentence in English. Therefore, assuming that our use of the term "true" is adequate, we can assert the equivalence (3) in which "p" is replaced by "s". Thus we can state:

> **(8) "s" is true if and only if s.**

We now recall that "s" stands for the whole sentence (6). Hence we can replace "s" by (6) on the right side of (8); we then obtain

> **(9) "s" is true if and only if the sentence printed in red on page 65 of the June 1969 issue of *Scientific American* is false.**

By now comparing (8) and (9), we conclude:

> **(10) "s" is false if and only if "s" is true.**

This leads to an obvious contradiction: "s" proves to be both true and false. Thus we are confronted with an antinomy. The above formulation of the antinomy of the liar is due to the Polish logician Jan Łukasiewicz.

Some more involved formulations of this antinomy are also known. Imagine, for instance, a book of 100 pages, with just one sentence printed on each page. [[66]]
On page 1 we read:

> **The sentence printed on page 2 of this book is true.**

On page 2 we read:

> **The sentence printed on page 3 of this book is true.**

And so it goes on up to page 99. However, on page 100, the last page of the book, we find:

> **The sentence printed on page 1 of this book is false.**

Assume that the sentence printed on page 1 is indeed false. By means of an argument which is not difficult but is very long and requires leafing through the entire book, we conclude that our assumption is wrong. Consequently we assume now that the sentence printed on page 1 is true—and, by an argument which is as easy and as long as the original one, we convince ourselves that the new assumption is wrong as well. Thus we are again confronted with an antinomy.

It turns out to be an easy matter to compose many other "antinomial books" that are variants of the one just described. Each of them has 100 pages. Every page contains just one sentence, and in fact a sentence of the form:

> **The sentence printed on page 00 of this book is XX.**

In each particular case "XX" is replaced by one of the words "*true*" or "*false*", while "00" is replaced by one of the numerals "1", "2",. . . , "100"; the same numeral may occur on many pages. Not every variant of the original book composed according to these rules actually yields an antinomy. The reader who is fond of logical puzzles will hardly find it difficult to describe all those variants that do the job. The following warning may prove useful in this connection. Imagine that somewhere in the book, say on page 1, it is said that the sentence on page 3 is true, while somewhere else, say on page 2, it is claimed that the same sentence is false. From this information it does not follow at all that our book is "antinomial"; we can only draw the conclusion that either the sentence on page 1 or the sentence on page 2 must be false. An antinomy does arise, however, whenever we are able to show that one of the sentences in the book is both true and false, independent of any assumptions concerning the truth or falsity of the remaining sentences.

The antinomy of the liar is of very old origin. It is usually ascribed to the Greek logician Eubulides; it tormented many ancient logicians and caused the premature death of at least one of them, Philetas of Cos. A number of other antinomies and paradoxes were found in antiquity, in the Middle Ages, and in modern times. Although many of them are now entirely forgotten, the antinomy of the liar is still analyzed and discussed in contemporary writings. Together with some recent antinomies discovered around the turn of the century (in particular, the antinomy of Russell), it has had a great impact on the development of modern logic.

Two diametrically opposed approaches to antinomies can be found in the literature of the subject. One approach is to disregard them, to treat them as sophistries, as jokes that are not serious but malicious, and that aim mainly at showing the cleverness of the man who formulates them. The opposite approach is characteristic of certain thinkers of the 19th century and is still represented, or was so a short while ago, in certain parts of our globe. According to this approach antinomies constitute a very essential element of

[b] [[Printed in red, and on page 65, in the original]]

human thought; they must appear again and again in intellectual activities, and their presence is the basic source of real progress. As often happens, the truth is probably somewhere in between. Personally, as a logician, I could not reconcile myself with antinomies as a permanent element of our system of knowledge. However, I am not the least inclined to treat antinomies lightly. The appearance of an antinomy is for me a symptom of disease. Starting with premises that seem intuitively obvious, using forms of reasoning that seem intuitively certain, an antinomy leads us to nonsense, a contradiction. Whenever this happens, we have to submit our ways of thinking to a thorough revision, to reject some premises in which we believed or to improve some forms of argument which we used. We do this with the hope not only that the old antinomy will be disposed of but also that no new one will appear. To this end we test our reformed system of thinking by all available means, and, first of all, we try to reconstruct the old antinomy in the new setting; this testing is a very important activity in the realm of speculative thought, akin to carrying out crucial experiments in empirical science.

From this point of view consider now specifically the antinomy of the liar. The antinomy involves the notion of truth in reference to arbitrary sentences of common English; it could easily be reformulated so as to apply to other natural languages. We are confronted with a serious problem: how can we avoid the contradictions induced by this antinomy? A radical solution of the problem which may readily occur to us would be simply to remove the word "true" from the English vocabulary or at least to abstain from using it in any serious discussion.

Those people to whom such an amputation of English seems highly unsatisfactory and illegitimate may be inclined to accept a somewhat more compromising solution, which consists in adopting what could be called (following the contemporary Polish philosopher Tadeusz Kotarbiński) "the nihilistic approach to the theory of truth". According to this approach, the word "true" has no independent meaning but can be used as a component of the two meaningful expressions "it is true that" and "it is not true that". These expressions are thus treated as if they were single words with no organic parts. The meaning ascribed to them is such that they can be immediately eliminated from any sentence in which they occur. For instance, instead of saying

it is true that all cats are black

we can simply say

all cats are black,

and instead of

it is not true that all cats are black

we can say

not all cats are black.

In other contexts the word "true" is meaningless. In particular, it cannot be used as a real predicate qualifying names of sentences. Employing the terminology of medieval logic, we can say that the word "true" can be used syncategorematically in some special situations, but it cannot ever be used categorematically.

To realize the implications of this approach, consider the sentence which was the starting point for the antinomy of the liar; that is, the sentence printed in red on page 65 in this magazine. From the "nihilistic" point of view it is not a meaningful sentence, and the antinomy simply vanishes. Unfortunately, many uses of the word "true", which otherwise seem quite legitimate and reasonable, are similarly affected by this approach. Imagine, for instance, that a certain term occurring repeatedly in the works [[67]] of an ancient mathematician admits of several interpretations. A historian of science who studies the works arrives at the conclusion that under one of these interpretations all the theorems stated by the mathematician prove to be true; this leads him naturally to the conjecture that the same will apply to any work of this mathematician that is not known at present but may be discovered in the future. If, however, the historian of science shares the "nihilistic" approach to the notion of truth, he lacks the possibility of expressing his conjecture in words. One could say that truth-theoretical "nihilism" pays lip service to some popular forms of human speech, while actually removing the notion of truth from the conceptual stock of the human mind.

We shall look, therefore, for another way out of our predicament. We shall try to find a solution that will keep the classical concept of truth essentially intact. The applicability of the notion of truth will have to undergo some restrictions, but the notion will remain available at least for the purpose of scholarly discourse.

To this end we have to analyze those features of the common language that are the real source of the antinomy of the liar. When carrying through this analysis, we notice at once an outstanding feature of this language—its all-comprehensive, universal character. The common language is universal and is intended to be so. It is supposed to provide adequate facilities for expressing everything that can be expressed at all, in any language whatsoever; it is continually expanding to satisfy this requirement. In particular, it is semantically universal in the following sense. Together with the linguistic objects, such as sentences and terms, which are components of this language, names of these objects are also included in the language (as we know, names of expressions can be obtained by putting the expressions in quotes); in addition, the language contains semantic terms such as "truth", "name", "designation", which directly or indirectly refer to the relationship between linguistic objects and what is expressed by them. Consequently, for every sentence formulated in the common language, we can form in the same language another sentence to the effect that the first sentence is true or that it is false. Using an additional "trick" we can even construct in the language what is sometimes called a self-referential sentence, that is, a sentence *S* which asserts the fact that *S* itself is true or that it is false. In case *S* asserts its own falsity we [[68]] can show by means of a simple argument that *S* is both true and false—and we are confronted again with the antinomy of the liar.

There is, however, no need to use universal languages in all possible situations. In particular, such languages are in general not needed for the purposes of science (and by science I mean here the whole realm of intellectual inquiry). In a particular branch of science, say in chemistry, one discusses certain special objects, such as elements, molecules, and so on, but not for instance linguistic objects such as sentences or terms. The language that is well adapted to this discussion is a restricted language with a limited vocabulary; it must contain names of chemical objects, terms such as "element" and "molecule", but not names of linguistic objects; hence it does not have to be semantically universal. The same applies to most of the other branches of science. The situation becomes somewhat confused when we turn to linguistics. This is a science in which we study languages; thus the language of linguistics must certainly be provided with names of linguistic objects. However, we do not have to identify the language of linguistics with the universal language or any of the languages that are objects of linguistic discussion, and we are not bound to assume that we use in linguistics one and the same language for all discussions. The language of linguistics has to contain the names of linguistic components of the languages discussed but not the names of its own components; thus, again, it does not have to be semantically universal. The same applies to the language of logic, or rather of that part of logic known as metalogic and metamathematics; here we again concern ourselves with certain languages, primarily with languages of logical and mathematical theories (although we discuss these languages from a different point of view than in the case of linguistics).

The question now arises whether the notion of truth can be precisely defined, and thus a consistent and adequate usage of this notion can be established at least for the semantically restricted languages of scientific discourse. Under certain conditions the answer to this question proves to be affirmative. The main conditions imposed on the language are that its full vocabulary should be available and its syntactical rules concerning the formation of sentences and other meaningful expressions from words listed in the vocabulary should be precisely formulated. Furthermore, the syntactical rules should be purely formal, that is, they should refer exclusively to the form (the shape) of expressions; the function and the meaning of an expression should depend exclusively on its form. In particular, looking at an expression, one should be able in each case to decide whether or not the expression is a sentence. It should never happen that an expression functions as a sentence at one place while an expression of the same form does not function so at some other place, or that a sentence can be asserted in one context while a sentence of the same form can be denied in another. (Hence it follows, in particular, that demonstrative pronouns and adverbs such as "this" and "here" should not occur in the vocabulary of the language.) Languages that satisfy these conditions are referred to as formalized languages. When discussing a formalized language there is no need to distinguish between expressions of the same form which have been written or uttered in different places; one often speaks of them as if they were one and the same expression. The reader may have noticed we sometimes use this way of speaking even when discussing a natural language, that is, one which is not formalized; we do so for the sake of simplicity, and only in those cases in which there seems to be no danger of confusion.

Formalized languages are fully adequate for the presentation of logical and mathematical theories; I see no essential reasons why they cannot be adapted for use in other scientific disciplines and in particular to the development of theoretical parts of empirical sciences. I should like to emphasize that, when using the term "formalized languages", I do not refer exclusively to linguistic systems that are formulated entirely in symbols, and I do not have in mind anything essentially opposed to natural languages. On the contrary, the only formalized languages that seem to be of real interest are those which are fragments of natural languages (fragments provided with complete vocabularies and precise syntactical rules) or those which can at least be adequately translated into natural languages.

There are some further conditions on which the realization of our program depends. We should make a strict distinction between the language which is the object of our discussion and for which in particular we intend to construct the definition of truth, and the language in which the definition is to be formulated and its implications are to be studied. The latter is referred to as the metalanguage and the former as the object-language. The metalanguage must be sufficiently rich; in particular, it must include the object-language as a part. In fact, according to our stipulations, an adequate definition of truth will imply as consequences all partial definitions of this notion, that is, all equivalences of form (3):

"*p*" **is true if and only if** *p*,

where "*p*" is to be replaced (on both sides of the equivalence) by an arbitrary sentence of the object-language. Since all these consequences are formulated in the metalanguage, we conclude that every sentence of the object-language must also be a sentence of the metalanguage. Furthermore, the metalanguage must contain names for sentences (and other expressions) of the object-language, since these names occur on the left sides of the above equivalences. It must also contain some further terms that are needed for the discussion of the object-language, in fact terms denoting certain special sets of expressions, relations between expressions, and operations on expressions; for instance, we must be able to speak of the set of all sentences or of the operation of juxtaposition, by means of which, putting one of two given expressions immediately after the other, we form a new expression.

Finally, by defining truth, we show that semantic terms (expressing relations between sentences of the object-language and objects referred to by these sentences) can be introduced in the metalanguage by means of definitions. Hence we conclude that the metalanguage which provides sufficient means for defining truth must be essentially richer than the object-language; it cannot coincide with or be translatable into the latter, since otherwise both languages would turn out to be semantically universal, and the antinomy of the liar could be reconstructed in both of them. We shall return to this question in the last section of this article.

If all the above conditions are satisfied, the construction of the desired definition of truth presents no essential difficulties. Technically, however, it is too involved to be explained here in detail. For any given sentence of the object-language one can easily formulate the corresponding partial definition of form (3). Since, however, the set of all sentences in the object-language is as a rule infinite, whereas every sentence of the metalanguage is a finite string of signs, we cannot arrive at a general defi- [[69]] nition simply by forming the logical conjunction of all partial definitions. Nevertheless, what we eventually obtain is in some intuitive sense equivalent to the imaginary infinite conjunction. Very roughly speaking, we proceed as follows. First, we consider the simplest sentences, which do not include any other sentences as parts; for these simplest sentences we manage to define truth directly (using the same idea that leads to partial definitions). Then, making use of syntactical rules which concern the formation of more complicated sentences from simpler ones, we extend the definition to arbitrary compound sentences; we apply here the method known in mathematics as definition by recursion. (This is merely a rough approximation of the actual procedure. For some technical reasons the method of recursion is actually applied to define, not the notion of truth, but the related semantic notion of satisfaction. Truth is then easily defined in terms of satisfaction.)

On the basis of the definition thus constructed we can develop the entire theory of truth. In particular, we can derive from it, in addition to all equivalences of form (3), some consequences of a general nature, such as the famous laws of contradiction and of excluded middle. By the first of these laws, no two sentences one of which is the negation of the other can both be true; by the second law, no two such sentences can both be false.

The Notion of Proof

Whatever may be achieved by constructing an adequate definition of truth for a scientific language, one fact seems to be certain: the definition does not carry with it a workable criterion for deciding whether particular sentences in this language are true or false (and indeed it is not designed at all for this purpose). Consider, for example, a sentence in the language of elementary high school geometry, say "the three bisectors of every triangle meet in one point". If we are interested in the question whether this sentence is true and we turn to the definition of truth for an answer, we are in for a disappointment. The only bit of information we get is that the sentence is true if the three bisectors of a triangle always meet in one point, and is false if they do not always meet; but only a geometrical inquiry may enable us to decide which is actually the case. Analogous remarks apply to sentences from the domain of any other particular science: to [[70]] decide whether or not any such sentence is true is a task of the science itself, and not of logic or the theory of truth.

Some philosophers and methodologists of science are inclined to reject every definition that does not provide a criterion for deciding whether any given particular object falls under the notion defined or not. In the methodology of empirical sciences such a tendency is represented by the doctrine of operationalism; philosophers of mathematics who belong to the constructivist school seem to exhibit a similar tendency. In both cases, however, the people who hold this opinion appear to be in a small minority. A consistent attempt to carry out the program in practice (that is, to develop a science without using undesirable definitions) has hardly ever been made. It seems clear that under this program much of contemporary mathematics would disappear, and theoretical parts of physics, chemistry, biology, and other empirical sciences would be severely mutilated. The definitions of such notions as atom or gene as well as most definitions in mathematics do not carry with them any criteria for deciding whether or not an object falls under the term that has been defined.

Since the definition of truth does not provide us with any such criterion and at the same time the search for truth is rightly considered the essence of scientific activities, it appears as an important problem to find at least partial criteria of truth and to develop procedures that may enable us to ascertain or negate the truth (or at least the likelihood of truth) of as many sentences as possible. Such procedures are known indeed; some of them are used exclusively in empirical science and some primarily in deductive science. The notion of proof—the second notion to be discussed in this paper—refers just to a procedure of ascertaining the truth of sentences which is employed primarily in deductive science. This procedure is an essential element of what is known as the axiomatic method, the only method now used to develop mathematical disciplines.

The axiomatic method and the notion of proof within its framework are products of a long historical development. Some rough knowledge of this development is probably essential for the understanding of the contemporary notion of proof.

Originally a mathematical discipline was an aggregate of sentences that concerned a certain class of objects or phenomena, were formulated by means of a certain stock of terms, and were accepted as true. This aggregate of sentences lacked any structural order. A sentence was accepted as true either because it seemed intuitively evident, or else because it was proved on the basis of some intuitively evident sentences, and thus was shown, by means of an intuitively certain argument, to be a consequence of these other sentences. The criterion of intuitive evidence (and intuitive certainty of arguments) was applied without any restrictions; every sentence recognized as true by means of this criterion was automatically included in the discipline. This description seems to fit, for instance, the science of geometry as it was known to ancient Egyptians and Greeks in its early, pre-Euclidean stage.

It was realized rather soon, however, that the criterion of intuitive evidence is far from being infallible, has no objective character, and often leads to serius errors. The entire subsequent development of the axiomatic method can be viewed as an expression of the tendency to restrict the recourse to intuitive evidence.

This tendency first revealed itself in the effort to prove as many sentences as possible, and hence to restrict as much as possible the number of sentences accepted as true merely on the basis of intuitive evidence. The ideal from this point of view would be to prove every sentence that is to be accepted as true. For obvious reasons this ideal cannot be realized. Indeed, we prove each sentence on the basis of other sentences, we prove these other sentences on the basis of some further sentences, and so on: if we are to avoid both a vicious circle and an infinite regress, the procedure must be discontinued somewhere. As a compromise between that unattainable ideal and the realizable possibilities, two principles emerged and were subsequently applied in constructing mathematical disciplines. By the first of these principles every discipline begins with a list of a small number of sentences, called axioms or primitive sentences, which seem to be intuitively evident and which are recognized as true without any further justification. According to the second principle, no other sentence is accepted in the discipline as true unless we are able to prove it with the exclusive help of axioms and those sentences that were previously proved. All the sentences that can be recognized as true by virtue of these two principles are called theorems, or provable sentences, of the given discipline. Two analogous principles concern the use of terms in constructing the discipline. By the first of them we list at the beginning a few terms, called undefined or primitive terms, which appear to be directly understandable and which we decide to use (in formulating and proving theorems) without explaining their meanings; by the second principle we agree not to use any further term unless we are able to explain its meaning by defining it with the help of undefined terms and terms previously defined. These four principles are cornerstones of the axiomatic method; theories developed in accordance with these principles are called axiomatic theories.

As is well known, the axiomatic method was applied to the development of geometry in the *Elements* of Euclid about 300 B.C. Thereafter it was used for over 2,000 years with practically no change in its main principles (which, by the way, were not even explicitly formulated for a long period of time) nor in the general approach to the subject. However, in the 19th and 20th centuries the concept of the axiomatic method did undergo a profound evolution. Those features of the evolution which concern the notion of proof are particularly significant for our discussion.

Until the last years of the 19th century the notion of proof was primarily of a psychological character. A proof was an intellectual activity that aimed at convincing oneself and others of the truth of a sentence discussed; more specifically, in developing a mathematical theory proofs were used to convince ourselves and others that a sentence discussed had to be accepted as true once some other sentences had been previously accepted as such. No restrictions were put on arguments used in proofs, except that they had to be intuitively convincing. At a certain period, however, a need began to be felt for submitting the notion of proof to a deeper analysis that would result in restricting the recourse to intuitive evidence in this context as well. This was probably related to some specific developments in mathematics, in particular to the discovery of non-Euclidean geometries. The analysis was carried out by logicians, beginning with the German logician Gottlob Frege; it led to the introduction of a new notion, that of a *formal proof*, which turned out to be an adequate substitute and an essential improvement over the old psychological notion.

The first step toward supplying a mathematical theory with the notion of [[75]] a formal proof is the formalization of the language of the theory, in the sense discussed previously in connection with the definition of truth. Thus formal syntactical rules are provided which in particular enable us simply by looking at shapes of expressions, to distinguish a sentence from an expression that is not a sentence. The next step consists in formulating a few rules of a different nature, the so-called rules of proof (or of inference). By these rules a sentence is regarded as directly derivable from given sentences if, generally speaking, its shape is related in a prescribed manner to the shapes of given sentences. The number of rules of proof is small, and their content is simple. Just like the syntactical rules, they all have a formal character, that is, they refer exclusively to shapes of sentences involved. Intuitively all the rules of derivation appear to be infallible, in the sense that a sentence which is directly derivable from true sentences by any of these rules must be true itself. Actually the infallibility of the rules of proof can be established on the basis of an adequate definition of truth. The best-known and most important example of a rule of proof is the rule of detachment known also as *modus ponens*. By this rule (which in some theories serves as the only rule of proof) a sentence "q" is directly derivable from two given sentences if one of them is the conditional sentence "if p, then q" while the other is "p"; here "p" and "q" are, as usual, abbreviations of any two sentences of our formalized language. We can now explain in what a formal proof of a given sentence consists. First, we apply the rules of proof to axioms and obtain new sentences that are directly derivable from axioms; next, we apply the same rules to new sentences, or jointly to new sentences and axioms, and obtain further sentences; and we continue this process. If after a finite number of steps we arrive at a given sentence, we say that the sentence has been formally proved. This can also be expressed more precisely in the following way: a formal proof of a given sentence consists in constructing a finite sequence of sentences such that (1) the first sentence in the sequence is an axiom, (2) each of the following sentences either is an axiom or is directly derivable from some of the sentences that precede it in the sequence, by virtue of one of the rules of proof, and (3) the last sentence in the sequence is the sentence to be proved. Changing somewhat the use of the term "proof", we can even say that a formal proof of a sentence is simply any finite sequence of sentences with the three properties just listed.

An axiomatic theory whose language has been formalized and for which the notion of a formal proof has been supplied is called a formalized theory. We stipulate that the only proofs which can be used in a formalized theory are formal proofs; no sentence can be accepted as a theorem unless it appears on the list of axioms or a formal proof can be found for it. The method of presenting a formalized theory at each stage of its development is in principle very elementary. We list first the axioms and then all the known theorems in such an order that every sentence on the list which is not an axiom can be directly recognized as a theorem, simply by comparing its shape with the shapes of sentences that precede it on the list; no complex processes of reasoning and convincing are involved. (I am not speaking here of psychological processes by means of which the theorems have actually been discovered.) The recourse to intuitive evidence has been indeed considerably restricted; doubts concerning the truth of theorems have not been entirely eliminated but have been reduced to possible doubts concerning the truth of the few sentences listed as axioms and the infallibility of the few simple rules of proof. It may be added that the process of introducing new terms in the language of a theory can also be formalized by supplying special formal rules of definitions.

It is now known that all the existing mathematical disciplines can be presented as formalized theories. Formal proofs can be provided for the deepest and most complicated mathematical theorems, which were originally established by intuitive arguments.

The Relationship of Truth and Proof

It was undoubtedly a great achievement of modern logic to have replaced the old psychological notion of proof, which could hardly ever be made clear and precise, by a new simple notion of a purely formal character. But the triumph of the formal method carried with it the germ of a future setback. As we shall see, the very simplicity of the new notion turned out to be its Achilles heel.

To assess the notion of formal proof we have to clarify its relation to the notion of truth. After all, the formal proof, just like the old intuitive proof, is a procedure aimed at acquiring new true sentences. Such a procedure will be adequate only if all sentences acquired with its help prove to be true and all true sentences can be acquired with its help. Hence the problem naturally arises: is the formal proof actually an adequate procedure for acquiring truth? In other words: does the set of all (formally) provable sentences coincide with the set of all true sentences?

To be specific, we refer this problem to a particular, very elementary mathematical discipline, namely to the arithmetic of natural numbers (the elementary number theory). We assume that this discipline has been presented as a formalized theory. The vocabulary of the theory is meager. It consists, in fact, of variables such as "m", "n", "p",... representing arbitrary natural numbers; of numerals "0", "1", "2", ... denoting particular numbers; of symbols denoting some familiar relations between numbers and operations on numbers such as "$=$", "$<$", "$+$", "$-$"; and, finally, of certain logical terms, namely sentential connectives ("and", "or", "if", "not") and quantifiers (expressions of the form "for every number m" and "for some number m"). The syntactical rules and the rules of proof are simple. When speaking of sentences in the subsequent discussion, we always have in mind sentences of the formalized language of arithmetic.

We know from the discussion of truth in the first section that, taking this language as the object-language, we can construct an appropriate metalanguage and formulate in it an adequate definition of truth. It proves convenient in this context to say that what we have thus defined is the set of true sentences; in fact, the definition of truth states that a certain condition formulated in the metalanguage is satisfied by all elements of this set (that is, all true sentences) and only by these elements. Even more readily we can define in the metalanguage the set of provable sentences; the definition conforms entirely with the explanation of the notion of formal proof that was given in the second section. Strictly speaking, the definitions of both truth and provability belong to a new theory formulated in the metalanguage and specifically designed for the study of our formalized arithmetic and its language. The new theory is called the metatheory or, more specifically, the meta-arithmetic. We shall not elaborate here on the way in which the metatheory is constructed—on its axioms, undefined terms, and so on. We only point out that it is within the framework of this metatheory [[76]] that we formulate and solve the problem of whether the set of provable sentences coincides with that of true sentences.

The solution of the problem proves to be negative. We shall give here a very rough account of the method by which the solution has been reached. The main idea is closely related to the one used by the contemporary American logician (of Austrian origin) Kurt Gödel in his famous paper on the incompleteness of arithmetic.

It was pointed out in the first section that the metalanguage which enables us to define and discuss the notion of truth must be rich. It contains the entire object-language as a part, and therefore we can speak in it of natural numbers, sets of numbers, relations among numbers, and so forth. But it also contains terms needed for the discussion of the object-language and its components; consequently we can speak in the metalanguage of expressions and in particular of sentences, of sets of sentences, of relations among sentences, and so forth. Hence in the metatheory we can study properties of these various kinds of objects and establish connections between them.

In particular, using the description of sentences provided by the syntactical rules of the object-language, it is easy to arrange all sentences (from the simplest ones through the more and more complex) in an infinite sequence and to number them consecutively. We thus correlate with every sentence a natural number in such a way that two numbers correlated with two different sentences are always different; in other words, we establish a one-to-one correspondence between sentences and numbers. This in turn leads to a similar correspondence between sets of sentences and sets of numbers, or relations among sentences and relations among numbers. In particular, we can consider numbers of provable sentences and numbers of true sentences. we call them briefly provable° numbers and true° numbers. Our main problem is reduced then to the question: are the set of provable° numbers and the set of true° numbers identical?

To answer this question negatively, it suffices, of course, to indicate a single property that applies to one set but not to the other. The property we shall actually exhibit may seem rather unexpected, a kind of *deus ex machina*.

The intrinsic simplicity of the notions of formal proof and formal provability will play a basic role here. We have seen in the second section that the meaning of these notions is explained essentially in terms of certain simple relations among sentences prescribed by a few rules of proof; the reader may recall here the rule of *modus ponens*. The corresponding relations among numbers of sentences are equally simple; it turns out that they can be characterized in terms of the simplest arithmetical operations and relations, such as addition, multiplication, and equality—thus in terms occurring in our arithmetical theory. As a consequence the set of provable° numbers can also be characterized in such terms. One can describe briefly what has been achieved by saying that the definition of provability has been translated from the metalanguage into the object-language.

On the other hand, the discussion of the notion of truth in common languages strongly suggests the conjecture that no such translation can be obtained for the definition of truth; otherwise the object-language would prove to be in a sense semantically universal, and a reappearance of the antinomy of the liar would be imminent. We confirm this conjecture by showing that, if the set of true° numbers could be defined in the language of arithmetic, the antinomy of the liar could actually be reconstructed in this language. Since, however, we are dealing now with a restricted formalized language, the antinomy would assume a more involved and sophisticated form. In particular, no expressions with an empirical content such as “the sentence printed in such-and-such place”, which played an essential part in the original formulation of the antinomy, would occur in the new formulation. We shall not go into any further details here.

Thus the set of provable° numbers does not coincide with the set of true° numbers, since the former is definable in the language of arithmetic while the latter is not. Consequently the sets of provable sentences and true sentences do not coincide either. On the other hand, using the definition of truth we easily show that all the axioms of arithmetic are true and all the rules of proof are infallible. Hence all the provable sentences are true; therefore the converse cannot hold. Thus our final conclusion is: there are sentences formulated in the language of arithmetic that are true but cannot be proved on the basis of the axioms and rules of proof accepted in arithmetic.

One might think that the conclusion essentially depends on specific axioms and rules of inference, chosen for our arithmetical theory, and that the final outcome of the discussion could be different if we appropriately enriched the theory by adjoining new axioms or new rules of inference. A closer analysis shows, however, that the argument depends very little on specific properties of the theory discussed, and that it actually extends to most other formalized theories. Assuming that a theory includes the arithmetic of natural numbers as a part (or that, at least, arithmetic can be reconstructed in it), we can repeat the essential portion of our argument in a practically unchanged form; we thus conclude again that the set of provable sentences of the theory is different from the set of its true sentences. If, moreover, we can show (as is frequently the case) that all the axioms of the theory are true and all the rules of inference are infallible, we further conclude that there are true sentences of the theory which are not provable. Apart from some fragmentary theories with restricted means of expression, the assumption concerning the relation of the theory to the arithmetic of natural numbers is generally satisfied, and hence our conclusions have a nearly universal character. (Regarding those fragmentary theories which do not include the arithmetic of natural numbers, their languages may not be provided with sufficient means for defining the notion of provability, and their provable sentences may in fact coincide with their true sentences. Elementary geometry and elementary algebra of real numbers are the best known, and perhaps most important, examples of theories in which these notions coincide.)

The dominant part played in the whole argument by the antinomy of the liar throws some interesting light on our earlier remarks concerning the role of antinomics in the history of human thought. The antinomy of the liar first appeared in our discussion as a kind of evil force with a great destructive power. It compelled us to abandon all attempts at clarifying the notion of truth for natural languages. We had to restrict our endeavors to formalized languages of scientific discourse. As a safeguard against a possible reappearance of the antinomy, we had to complicate considerably the discussion by distinguishing between a language and its metalanguage. Subsequently, however, in the new, restricted setting, we have managed to tame the destructive energy and harness it to peaceful, constructive purposes. The antinomy has not reappeared, but its basic idea has been used [[77]] to establish a significant metalogical result with far-reaching implications.

Nothing is detracted from the significance of this result by the fact that its philosophical implications are essentially negative in character. The result shows indeed that in no domain of mathematics is the notion of provability a perfect substitute for the notion of truth. The belief that formal proof can serve as an adequate instrument for establishing truth of all mathematical statements has proved to be unfounded. The original triumph of formal methods has been followed by a serious setback.

Whatever can he said to conclude this discussion is bound to be an anticlimax. The notion of truth for formalized theories can now be introduced by means of a precise and adequate definition. It can therefore be used without any restrictions and reservations in metalogical discussion. It has actually become a basic metalogical notion involved in important problems and results. On the other hand, the notion of proof has not lost its significance either. Proof is still the only method used to ascertain the truth of sentences within any specific mathematical theory. We are now aware of the fact, however, that there are sentences formulated in the language of the theory which are true but not provable, and we cannot discount the possibility that some such sentences occur among those in which we are interested and which we attempt to prove. Hence in some situations we may wish to explore the possibility of widening the set of provable sentences. To this end we enrich the given theory by including new sentences in its axiom system or by providing it with new rules of proof. In doing so we use the notion of truth as a guide; for we do not wish to add a new axiom or a new rule of proof if we have reason to believe that the new axiom is not a true sentence, or that the new rule of proof when applied to true sentences may yield a false sentence. The process of extending a theory may of course be repeated arbitrarily many times. The notion of a true sentence functions thus as an ideal limit which can never be reached but which we try to approximate by gradually widening the set of provable sentences. (It seems likely, although for different reasons, that the notion of truth plays an analogous role in the realm of empirical knowledge.) There is no conflict between the notions of truth and proof in the development of mathematics; the two notions are not at war but live in peaceful coexistence.