9

Alfred Tarski (1901–1983), Alonzo Church (1903–1995), and Kurt Gödel (1906–1978)

C. ANTHONY ANDERSON

Alfred Tarski

Tarski, born in Poland, received his doctorate at the University of Warsaw under Stanislaw Lesniewski. In 1942, he was given a position with the Department of Mathematics at the University of California at Berkeley, where he taught until 1968.

Undoubtedly Tarski's most important philosophical contribution is his famous "semantical" definition of truth. Traditional attempts to define truth did not use this terminology and it is not easy to give a precise characterization of the idea. The underlying conception is that semantics concerns *meaning* as a relation between a linguistic expression and what it expresses, represents, or stands for. Thus "denotes," "designates," "names," and "refers to" are semantical terms, as is "expresses." The term "satisfies" is less familiar but also plausibly belongs in this category. For example, the number 2 is said to *satisfy* the equation "$x^2 = 4$," and by analogy we might say that Aristotle satisfies (or satisfied) the formula "x is a student of Plato."

It is not quite obvious that there is a meaning of "true" which makes it a semantical term. If we think of truth as a property of sentences, as distinguished from the more traditional conception of it as a property of beliefs or propositions, it turns out to be closely related to satisfaction. In fact, Tarski found that he could define truth in this sense in terms of satisfaction.

The goal which Tarski set himself (Tarski 1944, Woodger 1956) was to find a "materially adequate" and formally correct definition of the concept of truth as it applies to sentences. To be materially adequate a definition must "catch hold of the actual meaning of an old notion," rather than merely "specify[ing] the meaning of a familiar word used to denote a novel notion" (Woodger 1956: 341). Again, in discussing the material adequacy of some of his other definitions, Tarski writes, "Now the question arises of whether *the definitions just constructed* (the formal rigour of which raises no objection) *are also adequate materially*; in other words *do they in fact grasp the current meaning of the notion as it is known intuitively?*" (Woodger 1956: 128–9).

To determine whether or not a proposed definition of a certain concept is materially adequate, Tarksi thinks that we must first formulate a *criterion* of material adequacy for such a definition: a precise condition which the definition must meet and which will guarantee that the defined notion is faithful to the original intuitive conception. Of

course, whether a proposed condition really guarantees sufficient conformity to the old notion is subject to critical review.

The requirement of *formal correctness* means that the proposed definition must be non-circular and that it must meet other logical constraints on acceptable definitions. One of the traditional requirements is that a definition must not define something in terms of things which are less clear than it. Tarksi even maintains that it must be specified which previously adopted terms are to be used in giving the definition and requires that the formal structure of the language in which the definition is to be given be precisely described.

These are rigorous constraints. The motivating idea seems to be that only under such conditions can we hope to *prove* the material adequacy and formal correctness of a definition of truth.

Tarski proposes as a criterion of material adequacy for a definition of truth that the definition shall have as logical consequences all instances of Schema (T):

(T) X is true if and only if p,

where "X" is replaced by a name of an arbitrary sentence of the language in question and "p" is replaced by that very sentence (or by a sentence with exactly the same meaning). The name in question must be a quotation-mark name or at least a name which necessarily designates the sentence. An appropriate instance of Schema (T) is thus such a thing as:

(S) "Snow is white" is true if and only if snow is white.

On the left-hand side of this "if and only if" there occurs a name of a certain sentence – which name is constructed by enclosing the sentence in question in quotation marks. Then using that name to *mention* the sentence, the property of being true is predicated of the sentence. On the right-hand side of the equivalence, the very sentence, which is named on the left and said there to be true, is *used*. The thing may appear to be a triviality and perhaps that is all to the good. The condition, after all, is supposed to constrain an adequate definition in such a way that satisfying this condition guarantees that the definition catches hold of the actual meaning of the term "true."

Note carefully that Schema (T) is *not* Tarski's definition of truth. That a definition should imply all instances of Schema (T) is the criterion of adequacy for the definition. But Tarski does seem to think that all the instances of (T) together completely capture the meaning of "true." If we could form an infinite conjunction, connecting all the instances with "and," we would have a complete specification of the semantical conception of truth. This is not an acceptable procedure according to the usual rules of definition, but a correct definition would be obtained if we could somehow achieve the same effect.

Now the conditions which have already been given for an acceptable definition of truth require that the language involved be specified quite precisely. Natural languages do not have, or at least we do not know, rules which determine exactly what its expressions are; for example, the sentences of English are not precisely specified. If we ignore this and set as our task to give a definition of truth for a natural language, say English, we encounter a paradox. No predicate of a sufficiently expressive language such as English can have the property that it validates every instance of Schema (T). And this

is so whether the predicate is defined or not. The proof of this appeals to the infamous liar antinomy (or paradox). In a very simple version the antinomy (something "contrary to law") goes like this. Consider

(A) A is not true.

That is, consider the sentence "A is not true," which sentence we have decided to name "A." Now Schema (T) implies:

(1) "A is not true" is true if and only if A is not true.

But observing that the sentence A is the very sentence "A is not true", we may assert:

(2) A = "A is not true."

If two things are identical, then they share all the same properties. So, substituting the left-hand side of (2) for the right-hand side in (1), we get:

(3) A is true if and only if A is not true.

In the propositional calculus, this has the form:

(4) $P \equiv \sim P$,

"P if and only if not-P" and this is equivalent to the explicit contradiction:

(5) $P \,\&\, \sim P$,

"P and not-P."

Something must give. If we are unwilling to give up the usual laws of logic, since (2) is undeniable, it appears that we must alter or modify Schema (T), our criterion allegedly determined by the very meaning of "true."

Tarski concludes, somewhat hastily, that ordinary language is inconsistent. The concept of truth must conform to Schema (T), but if we have such sentences as A, we arrive at a contradiction. The problem, says Tarski, is that natural languages are *semantically closed*, that is, they contain within themselves the terms and machinery for doing their own semantics. For example "is true in English" is itself a predicate of English. We must, he says, give our definition of truth in a *metalanguage* for the language whose sentences are in question. A metalanguage is a language which we may use to talk about another language. For example, in a book written in English which explains the grammar and meaning of the German language, the metalanguage is English. The language being studied is called the *object language*: in the case of this example, German. Further, claims Tarski, we must confine our attention to formalized languages which, unlike natural language, need not be semantically closed and which are otherwise precisely specified.

With these provisos, Tarski proceeds to show that definitions of truth can be given for object languages which do not contain semantical terms. His method of definition has the striking quality that the definition, given in a metalanguage, *does not itself use any semantical terms*. Because of the liar antinomy and other conundrums involving semantical notions, Tarski considered it important to give the definition in such a way that *no* semantical terms are presupposed as primitive or understood without definition.

To see how the definition would be given for a very simple formalized language, let us suppose that we have just two predicates: "R," meaning *is red*, and "S," meaning *is square*. In addition, suppose that the language contains a variable x; a sign for negation, say "–"; for conjunction, "&"; and a notation for universal quantification, "∀" meaning *For every*. Thus, for example, we can write "$\forall x - S(x)$" for "For every object x, x is not square" or, more naturally, "Nothing is square."

We assume that our metalanguage contains the means of expressing at least the very same notions as the object language. Here we are using a bit of English as metalanguage so that we have the words "is red" to mean the same as the predicate "R" in the simple formalized language. Now let some domain of objects be selected as the collection of things we will be talking about. One can then define *satisfaction* for the object language:

(1) An object satisfies "$R(x)$" if and only if it is red.
(2) An object satisfies "$S(x)$" if and only if it is square.
(3) An object satisfies a negation $\ulcorner -\phi \urcorner$ if and only if it does not satisfy ϕ.
(4) An object satisfies a conjunction of the form $\ulcorner \phi \,\&\, \psi \urcorner$ if and only if it satisfies ϕ and it satisfies ψ.
(5) An object satisfies a universal quantification $\ulcorner \forall \times \phi \urcorner$ if and only if every object (in the domain) satisfies ϕ.

Here ϕ and ψ are *formulae* of the formalized language. These are expressions which we have not really defined but which include such things as "$R(x)$" ("x is red"), "$-[S(x)$ & $R(x)]$" ("It is not the case that x is square and x is red"), as well as sentences such as "$-\forall x - S(x)$" ("It is not the case that for every x, x is non-square," i.e. "Something is square").

This doesn't look like a definition, but in fact it really does completely explain the meaning of "satisfies" as it applies to our simple language. Using these definitional rules on complicated expressions we can proceed step by step to simpler expressions until we get down to cases covered by (1) and (2). And it may look as if we have some kind of vicious circularity. For example, we have used "and" (in the metalanguage) to define satisfaction for expressions (of the object language) containing "&." But the appearance is deceptive. We have assumed that whatever we can say in the object language, we can say in the metalanguage, but not necessarily vice versa. This assumption does not introduce any logical or philosophical difficulty into the definition.

Finally we define truth:

A sentence ϕ is true if and only if every object (in the domain) satisfies it. Again, we haven't really defined the sentences of our object language, but they will be expressions in which no occurrences of the variable are "dangling." For example, "$-\forall x - [R(x)$ & $S(x)]$" ("Something is red and square") is a sentence, as opposed to a formula such as "$R(x)$" ("x is red"). Here the variable x is just a placeholder, indefinitely indicating something or other, but no definite thing.

It is not obvious that this definition actually conforms to the criterion of material adequacy. But it does. It can be proved that every instance of Schema (T), confined to sentences of our object language, is a consequence of this definition. The whole thing may seem trivial, but it is really quite amazing that in an appropriate metalanguage truth can be defined without appealing to any semantical notions. This means that it

has been defined in terms of things which are clearer: they are just the concepts of logic together with the concepts of the object language.

It remains to mention Tarski's work on the notion of *logical consequence* (Woodger 1956: 409–20). This, like the notion of truth, was used in an intuitive way by logicians and philosophers before Tarski, but it was the latter who made the notion precise.

Consider once again our simple formalized language. Do not select a particular domain and particular meanings for "S" and "R." Rather, contemplate any arbitrary *interpretation* of the language – any domain of objects whatsoever and any appropriate meanings for these symbols. The logical symbols "–," "&," and so on, are to retain their original meanings throughout.

For any such specification, we can explain *truth under that interpretation* along the lines used above for the particular interpretation we were considering. Suppose that in some interpretation a particular sentence, say "$\forall x[\mathrm{S}(x) \,\&\, \mathrm{R}(x)]$," comes out true. Then in that interpretation certain other sentences will come out true as well. For example, "$\forall x\mathrm{S}(x)$" and "$\forall x\mathrm{R}(x)$." In fact, this will always happen. If an interpretation makes our example sentence true, that interpretation will also make these two sentences true. In such a case, Tarski says that the latter two sentences are *logical consequences* of the first sentence. In general, a sentence ψ is a logical consequence of a sentence ϕ if and only if every interpretation which makes ϕ come out true also makes ψ come out true. And a sentence is defined to be a logical consequence of a collection, or set, of sentences if every interpretation which makes every sentence in the set come out true also makes the sentence in question come out true. Finally, Tarski defines a sentence to be *logically valid* if it comes out true under every interpretation.

The importance of such a definition is that we can now strictly define what it is for something to be a valid argument in our language. And, of course, the study of valid arguments is at the very heart of the discipline of logic. Using these definitions we can then prove that certain systems of logical rules are "complete" in the sense of being adequate to their intended purpose of capturing all valid inferences expressible in the language. For example, certain formulations of first-order logic, the logic of such notions as *and, not, or, if, . . . then, not, some all*, and the like were proved complete by Kurt Gödel, to be discussed below.

These two things, his definition of the concept of truth for formalized languages and his explication of the concept of logical consequence are Tarski's distinctive philosophical contributions. They are substantial indeed.

Alonzo Church

In 1927, Church received his Ph.D. from Princeton, where he taught from 1929 to 1967. Thereafter, he taught at UCLA until 1990. He was a long-time editor of the *Journal of Symbolic Logic*, which he helped to found. Church's philosophical contributions largely concern questions about the foundations of logic and mathematics, especially their ontology, and topics in the philosophy of language and in the related area of intensional logic.

Church's thesis is a hypothesis concerning the identification of the mechanically computable or calculable functions discussed below in connection with Gödel's incompleteness theorem. Church proposed as a precise mathematical analysis of the idea of

such functions that they be identified with the *lambda-definable* functions. This latter notion is too technical to be explained in detail here. Alan Turing independently proposed the identification of the mechanically computable functions with functions computable in principle by a precisely definable sort of abstract "machine," now called a *Turing machine*. This identification turned out to be equivalent to Church's thesis. That is, the class of lambda-definable functions is exactly the same as the class of functions computable by a Turing machine. Other attempts to analyze the notion in question have always led to the same class of functions. The identification of the class of mechanically ("algorithmically") calculable functions with the class of lambda-definable or Turing machine-computable functions (the "Church–Turing thesis") is now almost universally accepted.

Church's theorem, to be carefully distinguished from Church's thesis, is a theorem of mathematical logic to the effect that there is no effective (= mechanical) procedure for deciding whether or not a formula of first-order logic is valid.

Church was a Platonist or, as he preferred, a realist about the entities apparently described and studied by mathematics and logic. Numbers and other mathematical entities are, he believed, objectively existing, mind-independent objects and mathematics itself consists of truths about these things. Logic seems to require, if formulated in full generality, *propositions*, *properties*, and "*individual concepts*." These kinds of things, usually called *intensional entities*, are supposed to be abstract, real, and objective entities suitable to be the meanings of expressions in various languages. Propositions, for example, are claimed to be the meanings of declarative sentences, the same for synonymous sentences, whether in a single language or in two or more different languages.

Church's general methodological viewpoint about the formal sciences was a kind of "hypothetico-deductive rationalism." According to this view, intuitions or feelings of self-evidence provide initial support for assumptions about abstract entities. The theories of these are to be *formalized*, stated using the precise language and terminology of symbolic logic, and the results are to be evaluated using the sorts of criteria common to scientific procedures in general. One way we evaluate theories is by deducing consequences and thereby determining whether they are adequate to account for the data. In the formal sciences Church took the data to include the accepted facts of mathematics and logic.

Many of Church's philosophical contributions appear in reviews in the *Journal of Symbolic Logic*. His relatively few papers devoted explicitly to philosophical topics usually concerned questions about meaning and related topics in the philosophy of language. There are also arguments against nominalism as it is sometimes espoused in connection with mathematics, logic, or semantics.

As a sample of the latter (Church 1950), consider a nominalist attempt to give an analysis of certain statements apparently about propositions. Suppose it is claimed that such a sentence as (1) "Seneca said that man is a rational animal" is to be analyzed as: (2) "There is a language *S*′ such that Seneca wrote as a sentence of *S*′ words whose translation from *S*′ into English is 'Man is a rational animal'." This may already seem excessively complicated, but simpler attempts to analyze statements about assertion so that they concern such relatively concrete things as sentences are subject to easy refutation. To bring out clearly that (2) will not do as an analysis of (1), Church uses the "translation test," a procedure whose invention is usually attributed to C. H. Langford.

If we translate (1) into German, we get (1′) "Seneca hat gesagt, das der Mensch ein vernünftiges Tier sei". In translating (2) into German, note carefully that the word "English" must be translated as "Englisch" (not as "deutsch") and the quotation which forms part of (2) is to be translated as "Man is a rational animal" (not as "Der Mensch ist ein vernünftiges Tier"). This latter translation, call it (2′), certainly would not convey anything like the information which would be conveyed to a German speaker (who spoke no English) by (1′). Thus, argues Church, (1′) is not an acceptable analysis of (1). The basic idea of the objection, which can be seen even without appealing to translation, is that (1) does not say anything about any particular language (and so neither does its translation (1′)), whereas (2) makes specific reference to English.

A philosophical argument which has a quite surprising conclusion is given by Church (Church 1956: 24–5) as a more precise version of reasoning offered by Gottlob Frege. The conclusion of the argument is that sentences *denote* truth-values, true sentences denoting Truth (or The True) and false sentences denoting Falsehood (or The False)! Put like this, the thesis seems quite incredible, even unintelligible. Why suppose that sentences "denote" anything at all? And what, we may ask, are these alleged "objects," Truth and Falsehood? These are good questions, but the essential point of Church's argument (and Frege's before him) could be stated like this: the truth or falsity of a sentence is the only thing that stands to the sentence as the denotation of a (complex) name stands to its parts. To see this take such a sentence as (a) "Sir Walter Scott is the author of *Waverley*." If we replace "the author of *Waverley*" by an expression which denotes the same, "the man who wrote twenty-nine Waverley Novels altogether," we get a new sentence: (b) "Sir Walter Scott is the man who wrote twenty-nine Waverley novels altogether." If we are supposing that the "denotation" of a sentence, whatever it is, is unchanged if a denoting part is replaced by another with the same denotation, then this new sentence must have the same denotation as the original. Further, it is plausible (Church claims) that the sentence, (c) "The number, such that Sir Walter Scott wrote that many Waverley Novels altogether is twenty-nine," is so close in meaning to (b) as to have the same "denotation" (again, without yet assuming that we know what this is). But now let us replace the denoting expression "The number, such that Sir Walter Scott wrote that many Waverley Novels altogether" in (c) by an expression with the same denotation, namely; "The number of counties in Utah" (which is in fact twenty-nine). We then get a sentence which is supposed to have the same denotation as (c), (d) "The number of counties in Utah is twenty-nine" (again assuming that a sentence does not change its denotation if a denoting part is replaced by another with the same denotation).

Now compare our original sentence (a) "Sir Walter Scott is the author of *Waverley*" with (d) "The number of counties in Utah is twenty-nine." By the reasoning just explained, these two sentences must have the same "denotation." But the only meaning-relevant feature which they seem to have in common is that both are true. A little reflection on such examples points to the conclusion that the only thing that can be expected to remain invariant under such substitutions is the truth or falsity of the original sentence. So if "denotation" has an analog for sentences, it will have to be the *truth-values*, truth and falsity, which may be seen as mathematical abstractions. (Compare the mathematical abstraction of numbers, as objects, from collections or from concepts of collections.) The Church–Frege argument here may not be conclusive,

but the analogy uncovered is striking and it may well be a useful theoretical assumption for semantics that sentences "denote" truth-values. (See Anderson 1998 for further discussion.)

Church's most important philosophical ideas are contained in his work on the foundations of intensional logic (Church 1951, 1973, 1974). Philosophers and logicians contrast intension and extension, but it is by no means easy to give a clear characterization of these notions. In the case of sentences, Church would maintain that the sense, or intension, of the sentence is the proposition which it expresses and the denotation, as already explained, is the truth-value of the sentence. Logic as standardly taught in philosophy and mathematics departments makes no significant distinction between sentences with the same truth-value; arguments which turn on finer distinctions of meaning are simply not treated. Similar distinctions hold between the *set* of things of which a predicate is true, the extension of the predicate, and the property conveyed by the predicate, its intension. Again, a distinction between the meaning, strictly so-called, of an expression such as "The present president of the US" (its intension) and what it stands for, the actual person, is needed. Here we might say, again with Church, that the meaning of the expression in the strict sense is the *concept* that it expresses, its intension, but what it denotes or stands for, the person or, more generally, the object, is its extension.

So, as already explained, Church calls the proposition expressed by a sentence its *sense* and the truth-value that it stands for its *denotation*. Predicates have properties as their senses and sets as their denotations, and individual expressions (e.g. descriptive names) have certain concepts as their senses and what they stand for as their denotations. The relationship that holds between the sense of an expression and what it denotes let us call the *concept relation*, and symbolize it by the capital Greek letter Δ (delta). Then propositions are concepts of truth-values, properties are concepts of sets, and individual concepts are concepts of the individual things that the concepts characterize. Generalizing our terminology (as Church does), call anything that is capable of being the sense of some expression a *concept*.

The intensional logic that Church envisioned would have two kinds of intensional axioms: logical principles about Δ and principles that would specify the essential characteristics of propositions and other complex concepts. In connection with the latter, Church took it to be especially important to have axioms which give, or correspond to, criteria of identity for complex concepts. A criterion of identity in the present case is a principle that determines the identity or difference of the complex concepts expressed by different sentences (or predicates or descriptive names) in terms of some known relation between the sentences (complex expressions) themselves. An example would be the principle that two sentences express the same proposition if and only if they are logically equivalent; in our example, that is, they have the same truth-value necessarily, or on logical grounds alone.

We have already explained that a function of numbers is a correlation of a certain kind. Thus, *square* or *squaring* is said to be a function from numbers to numbers. In general, any correlation between the things in two collections is called a function. Generally, a function is just any conceivable correlation between the things in one collection and the things in another (or, possibly, the same) collection; it is allowed that two or more things in the first collection be correlated with the same thing in the second.

A name of a function has both a sense and (in general) a denotation. The sense is therefore a concept of the function denoted by that name. For example the expression "The squaring function" denotes the function that takes each number into its square and it has as its sense what is conveyed by "x^2." The combination of a name of a function with the name of something to which the function is applied (an *argument* of the function) will also have a sense: a complex sense involving the sense of the function name and the sense of the name of the entity to which the function is being applied.

The importance of this idea appears in the observation that *any* complex expression may be construed as being built up from a function expression, together with expressions for one or more arguments to which the function is applied.

Now let us write "$\Delta(X,Y)$" to mean that X is a concept of Y. The axioms which Church took to govern the delta-relation are:

(C1) For every X, Y, and Z, if $\Delta(X,Y)$ and $\Delta(X,Z)$, then $Y = Z$.

(C2) For every F and F_1, if $\Delta(F_1,F)$, then for every X and X_1, if $\Delta(X_1,X)$, then $\Delta(F_1X_1,FX)$.

(C3) For every F and F_1, if for every X and X_1, $\Delta(X,X_1)$ implies that $\Delta(F_1X_1,FX)$, then $\Delta(F_1,F)$.

(C1) says that anything which is a concept of something is a concept of exactly one thing. In (C2) and (C3), F is any function and FX is the result of applying that function to an argument X; that is, FX is the entity which is correlated with X by the function. Where F_1 and X_1 are concepts, we have just written "F_1X_1" for the complex concept that results when the concept F_1 is combined with the concept X_1. In these terms, (C2) amounts to the claim that if an expression denoting a function is combined with an expression denoting an argument (in some possible language), then the sense of the complex expression is the result of combining the sense of the function name with the sense of the argument name.

The proposed axiom (C3) is more problematic. To understand and accept it, one really must go along with a hypothesis that Church proposes to simplify the logic of the system. Church assumes that *a concept of a function can be taken to be a function from concepts to concepts*. This is fine for axiom (C2), which is then just understood in such a way that combining a concept of a function with a concept of an argument is nothing more than applying a certain kind of function to a certain kind of argument. But axiom (C3) is much bolder. It amounts to the claim that *any* function from concepts to concepts satisfying a certain condition is a concept of a certain function. It says: if a function applied to a concept of an argument always yields a concept of the output of some function applied to the argument thus concepted, then the function (from concepts to concepts) is a concept of the function from objects to objects.

This axiom leads to various difficulties, which cannot be explained here (see Anderson 1998). It is fair to say that even the basic principles of intensional logic, as Church conceived it, are still not settled.

Intensional principles of the second sort – those supposed to individuate complex concepts – are also still problematic. Church proposed three heuristic principles to guide the formulation of such axioms: (A) that logically equivalent expressions express the same concept, (B) that expressions that have exactly the same

syntactical structure and whose corresponding parts have the same meanings express the same concept, and (C) expressions that can be obtained from one another by applying the logical operation of lambda conversion express the same concept. Lambda conversion is a logically valid transformation of expressions, which we do not attempt to explain here.

The idea behind (A) works well if, but only if, one is dealing with reasoning involving no finer distinctions of meaning than are involved in arguments turning on *modality*: necessity, possibility, impossibility, and similar conceptions. Suggestion (C) appeals to the technical notion of lambda-conversion which is very difficult to motivate from a philosophical point of view. Hands down, the notion urged in (B) is the most promising. Church (who agreed with the assessment just offered) tried to implement this approach several times in his published work, but technical and logical difficulties still block the way of a satisfactory theory.

It is fair to say that this project to which Church contributed fundamental and important work, to establish a comprehensive and adequate general intensional logic, has not yet been completed. But his successful philosophical contributions are impressive indeed.

Kurt Gödel

Kurt Gödel received his doctorate in 1930 at the University of Vienna. He emigrated to the United States in 1940 and soon afterwards became a member of the Institute for Advanced Studies at Princeton, New Jersey, until his death. Gödel, like Tarski and Church, is best known as a logician. But his logical discoveries are of profound significance for parts of philosophy: the philosophy of logic and mathematics, epistemology, and (perhaps) the philosophy of mind. In addition, in later years Gödel concentrated on philosophical questions and made strikingly original suggestions as to their solution, including an improved version of Anselm's famous ontological argument for the existence of God, as elaborated by Leibniz.

Gödel's most famous discoveries are his two *incompleteness theorems* (Gödel 1931). Here we will give outlines of modernized version of his proofs. Suppose that the arithmetic of the natural numbers (0, 1, 2, and so on) is formulated as an axiomatic system. The language used is a precisely specified symbolic, or formalized, language with axioms stating the basic properties of the natural numbers and rules of inference stating which sentences may be correctly inferred from others. A sequence of sentences beginning with axioms and constructed by applying the rules of inference is said to be a *proof* of the last sentence in the sequence, which latter is said to be a *theorem* of the system.

Gödel observed that we can assign numbers (now called "gödel numbers") to the syntactical entities of the axiomatic system. That is, one can correlate numbers with symbols, with complex expressions, and even with sequences of expressions such as proofs. These numbers are assigned to symbols of the object language in the metalanguage. But these numbers are part of the subject matter of the object language theory itself.

This done, we have a sort of indirect way of talking about expressions and sequences of expressions of the formal theory within the theory itself. By talking about the gödel

numbers of expressions and sequences of expressions, we can simulate, or model, talk about the expressions and sequences.

Gödel then proved that if the formal system of arithmetic meets certain minimal conditions of adequacy, then the set of gödel numbers of the sentences provable in the system (the theorems of the system) can be defined within that system. The condition of minimal adequacy is that the system of arithmetic be capable of expressing certain functions of natural numbers. A function of natural numbers is just a correlation between numbers and numbers, or between pairs of numbers and numbers, or . . . and so on. Intuitively, the functions of natural numbers that must be expressible are those whose values for given arguments can be "effectively calculated": calculated by means of an algorithm or recipe, mechanically (and mindlessly) computed in a manner available to a computing machine.

Next, it can be proved that the set of gödel numbers of *true* sentences ("true" being defined in the manner of Tarski) is *not* definable in arithmetic. The proof uses an argument which parallels the reasoning of the liar antinomy (!) but which is logically unexceptionable. The conclusion is that such a system of arithmetic cannot contain or define its own truth predicate (as applied to gödel numbers as surrogates for sentences).

But if the set of gödel numbers of provable sentences *is* definable in arithmetic and the set of gödel numbers of true sentences (of arithmetic) is *not* definable in that system, then the two sets have to be different. Therefore, either some sentence provable in arithmetic is not true or some true sentence of arithmetic is not provable. We cannot accept the former, at least if we have chosen a system of axioms which we can see to be true of the natural numbers. We conclude that *some true sentence of arithmetic is not a theorem of arithmetic*. Any formal system of arithmetic meeting reasonable conditions of adequacy will be *incomplete*. This is essentially Gödel's first incompleteness theorem.

Gödel's actual proof did not proceed in the way we have described. Rather, he assumed not that the axioms of arithmetic are true (and that the rules of inference preserve truth) but only that arithmetic is *consistent*: it is not possible to derive an actual contradiction from the axioms using the rules of logic. (Really, he assumed that arithmetic is *omega-consistent*, a stronger assumption than consistency which we need not explain here.) Then Gödel showed how to construct, given that arithmetic is consistent, a particular sentence G which is such that neither G nor its negation –G ("not-G") is a theorem. The proof proceeds in such a way that we could, with sufficient patience and longevity, actually write down a true sentence of arithmetic which, if arithmetic is formally consistent, cannot be proved in arithmetic. And one can see that if the consistency of arithmetic is accepted, the sentence G is the true but unprovable one, as opposed to –G.

This sentence G involves just quite ordinary arithmetical concepts such as "plus" and "times" together with the usual logical concepts "and," "not," "some," "all," "equals," and so on. It is worth noticing that, contrary to various popular expositions, Gödel's original proof does not involve self-reference in any sense. The true but unprovable sentence G does not "say" that it, itself, is unprovable. It is a sentence entirely about natural numbers and their properties and relations. But it is a sentence that simulates such a self-referential sentence in the sense that it is true if and only if it is not provable.

Well, so what? It is natural to suggest that the axioms of arithmetic with which we began are just not adequate and that some new axioms must be added. It is as if Euclid's

geometry had been formulated without the Parallel Postulate. One would simply have to add it, or some equivalent. However, if you consider the details of Gödel's proof, it is evident that a system obtained by adding any finite number of axioms will still be subject to the proof – although the unprovable but true sentence will be different. Indeed, even adding infinitely many new axioms in any "effective" way does not evade the proof. We must conclude that nothing that would count as a formal system can contain all the truths of arithmetic. (If you think that to be a truth of arithmetic is to be provable in arithmetic, then this result will be quite difficult to comprehend. But one conclusion we can draw from these considerations is that this identification cannot be correct.)

This much is already quite startling. The goal of mathematicians since Euclid has been to specify certain basic truths of mathematics and to justify all others by deduction from these. Gödel proved that this goal is unattainable! No matter what formal (alias "axiomatic") system is proposed, there will be truths of arithmetic (the most basic part of mathematics) which the system cannot prove. Of course to show this with mathematical precision requires that we precisely define "axiomatic system" or "formal system," but this can be done in a way that is undeniably correct.

Gödel's second incompleteness theorem builds on the first. Using the technique of pseudo-self-reference mentioned above, one can find a sentence of any (minimally adequate) formal system of arithmetic which "says" that the system itself is consistent. Call this sentence "Consis." Now the proof of Gödel's first incompleteness theorem can be mimicked within arithmetic to produce a proof of the conditional sentence: "If Consis, then G." Suppose it were possible to prove within arithmetic the sentence Consis which mirrors the proposition that arithmetic is consistent. Then it would be possible to prove (by modus ponens) the Gödel sentence G. But we already know, from the first incompleteness theorem, that if arithmetic is consistent, G cannot be proved therein. If we suppose, as we are certainly entitled to do, that the theorems of arithmetic are true, then arithmetic is consistent. We conclude that the sentence which (in an indirect sense) expresses that arithmetic is consistent, cannot itself be proved in arithmetic. And, of course, the sentence "expressing" that arithmetic is *inconsistent* will not be a theorem either. It too, like the sentence G, is undecidable in arithmetic: can neither be proved nor refuted therein.

We have been at some pains to dispel the impression that Gödel's proofs literally involve self-reference. What then does, for example, the sentence "expressing" the consistency of arithmetic look like? Well, like the sentence G, it is written entirely in the language of arithmetic (involving "plus," "times," and "equals," for example). In fact it can be seen as expressing a certain mathematical claim about a *polynomial*. Let P(x,y) be a polynomial involving just the two indicated variables and integral coefficients. Then the statement "For every y, there is an x, such that P(x,y) = 0" may be true or false, depending on the details of the polynomial. Problems of this sort, as to whether or not such a statement is correct, are called "diophantine" problems. More generally, suppose we have n variables $x_1, x_2, \ldots, x_n$ and m variables $y_1, y_2, \ldots, y_m$, and a polynomial (with integral coefficients) involving these, say "$P(x_1, \ldots, x_n, y_1, \ldots, y_m)$." Then the statement "expressing" the consistency of arithmetic is of the form:

For every $y_1, \ldots, y_m$, there are $x_1, \ldots, x_n$, such that $P(x_1, \ldots, x_n, y_1, \ldots, y_m) = 0$.

This is a purely arithmetical statement and, one would have supposed, a claim that arithmetical techniques should be able to settle, yea or nay. But they cannot. (The details of these proofs are rather difficult to grasp. The best introduction to them is Smullyan 1957. For precise details, Boolos and Jeffrey 1989 is excellent.)

The philosophical import of these results is more controversial. We have already observed that a certain mathematical program is thereby proved impossible: that of deducing all of mathematics from an axiomatic basis. If we think of metaphysics as including all necessary truths, mathematics not excluded, then the goal, perhaps most closely associated with Spinoza, of an axiomatic development of metaphysics is thus also proved impossible to achieve.

Gödel's own philosophical speculations on the import of his theorems are most clearly articulated in his *Collected Works* (1995: 304–23). Let us confine our attention to arithmetic and speak of the true sentences thereof as "objective mathematics." Human beings, using our presently accepted arithmetical assumptions, can certainly prove some of these sentences. Gödel calls the mathematical truths that human beings are capable of demonstrating "subjective mathematics" (perhaps not the best choice of terminology since these are all objectively true and indeed knowable to be such). Some of these may require axioms about the numbers which are presently unknown, but which can in principle be seen to be evident by human mathematicians.

Now consider again Gödel's second incompleteness theorem. The theorems of a formal system of arithmetic comprise a set of sentences that could be mechanically generated, one after the other. According to Gödel's second incompleteness theorem, any such system will be unable to prove (generate) the arithmetical sentence that "expresses" its own consistency. Recall that this was a certain sentence expressing a diophantine arithmetical problem. Gödel draws the following conclusion from this:

> Either mathematics is incompletable in this sense, that its evident axioms can never be comprised in a finite rule, that is to say the human mind (even within the realm of pure mathematics) infinitely surpasses the powers of any finite machine, or else there exist absolutely unsolvable diophantine problems of the type specified. (1995: 310)

To say that the evident axioms (and rules of inference) can be "comprised in a finite rule" is equivalent to the possibility of the formulation of arithmetic as a whole as a formal system. And this latter amounts to the possibility of being generated in a machine-like fashion. To say that a diophantine problem is absolutely unsolvable means that neither the statement nor its negation will ever be a theorem of subjective mathematics. There is no doubt that this conclusion actually does follow from Gödel's proof, together with the mentioned analysis of a finite mechanical procedure. Gödel himself thinks, and argues, that the second option is incorrect – there are no absolutely unsolvable arithmetical problems – but his arguments for this conclusion are not airtight and may be reasonably doubted.

Gödel also had an improvement on Anselm's ontological argument for the existence of God, especially as it was developed by Leibniz. Leibniz observed that the one

thing that needs to be proved to complete Anselm's proof is that the existence of God is *possible*. A number of commentators on the argument, including Kant, have observed that all that is really established by Anselm is that *if* God exists, then He necessarily exists. Now, given the nature of the proof, we may further conclude that this conditional itself is necessary. So, we may conclude further, by a standard principle of modal logic, that if it is *possible* that God exists, then it is *possible* that it is necessary that He exists. (The standard principle of modal logic in question is this: if p necessarily implies q, then if p is possible, then q is possible.) If we further suppose that we can somehow prove that it is possible that God exists (given a certain definition of "God"), then it follows that it is necessary that it is possible that God exists. But according to one plausible system of modal logic (standardly called "S5"), it then follows that it is necessary that God exists!

But can we prove that it is possible that God exists? Gödel thought that we can and that a version of the ontological argument is then cogent. His argument for this, which bears some resemblance to Leibniz's argument for the same conclusion, uses the idea of a *positive* property. Gödel doesn't really say very clearly what this conception involves, but he remarks that it has two possible interpretations, "positive in the moral-aesthetic sense" and positive in the sense of involving only "pure attribution." A being is defined to be *God-like* if it has every positive property. Then a property is defined to be an *essence* of an entity x if x has that property and it entails every other property that x has. An entity *necessarily exists*, by definition, if every essence of it is necessarily exemplified. Finally, Gödel assumes the following "axioms" about these concepts:

1 A property is positive if and only if its negation is not positive.
2 Any property entailed by a positive property is itself positive.
3 The property of being God-like is positive.
4 If a property is positive, it is necessarily positive.
5 The property of necessarily existing is positive.

From these it follows that it is possible that a God-like being exists and, by essentially the argument explained earlier, that such a being therefore exists, and indeed necessarily so. There are some problems with the argument, not the least of which is the obscurity of the notion of a positive property. For an able discussion of the argument and its alleged defects, see Adams 1970.

Gödel also thought that Einstein's Theory of Relativity has implications for idealism, in particular that it supports some ideas of Immanuel Kant. He argues that there is considerable reason to believe that "time is unreal" (1951: 555–62). Essentially the argument is this: If time and change are something real, then there must be such a thing as an objective and absolute lapse of time. But the Theory of Relativity, a well-confirmed scientific theory, seems to deny that there is such an objective lapse. Gödel considers various objections and explains the relevance of a technical contribution of his to that theory. That work, by the way, seems to imply the possibility of "time travel"!

In sum, we may say that Gödel's main work in logic is of profound philosophical significance and that his other philosophical work certainly deserves further careful study.

Bibliography

Works by Church, Gödel, and Tarski

Church, A., 1950: "On Carnap's Analysis of Statements of Assertion and Belief," *Analysis* 10, pp. 97–9.
1951: "A Formulation of the Logic of Sense and Denotation," in *Structure, Method and Meaning*, ed. P. Henle, H. M. Kallen, and S. K. Langer, New York: Liberal Arts Press.
1956: *Introduction to Mathematical Logic*, vol. I, Princeton, NJ: Princeton University Press.
1973: "Outline of a Revised Formulation of the Logic of Sense and Denotation (Part I)," *Noûs* 7, pp. 24–33.
1974: "Outline of a Revised Formulation of the Logic of Sense and Denotation (Part II)," *Noûs* 8, pp. 135–56.
Gödel, K., 1931: "Über formal unentscheidbare Sätze der *Principia Mathematica* und verwandter Systeme I," *Monatshefte für Mathematik und Physik* 38, pp. 173–98.
1951: "A remark about the relationship between relativity theory and idealistic philosophy," in *Albert Einstein: Philosopher-Scientist*, ed. P. A. Schilpp, New York: Tudor Publishing Company, pp. 555–62.
1986: *Collected Works*, vol. I, ed. S. Feferman, J. Dawson, and S. Kleene, New York: Oxford University Press.
1995: *Collected Works*, vol. III, ed. S. Feferman and J. Dawson, New York: Oxford University Press.
Tarski, A., 1944: "The Semantic Conception of Truth and the Foundations of Semantics," *Philosophy and Phenomenological Research* 4, pp. 341–76.
1956a: "On Definable Sets of Real Numbers," in J. H. Woodger, *Logic Semantics, Metamathematics: Papers from 1923 to 1938 by Alfred Tarski*, Oxford: Clarendon Press, pp. 110–42.
1956b: "On the Concept of Logical Consequence," in Woodger, pp. 409–20.
1956c: "The Concept of Truth in Formalized Languages," in Woodger, pp. 152–278.

Works by other authors

Adams, R. M. (1995) "Introductory note to *1970," in *Kurt Gödel: Collected Works*, vol. III, ed. S. Feferman, et al., Oxford: Oxford University Press, pp. 388–402.
Anderson, C. A. (1998) "Alonzo Church's contributions to philosophy and intensional logic," *Bulletin of Symbolic Logic* 4, pp. 129–71.
Boolos, G. S. and Jeffrey, R. C. (1989) *Computability and Logic*, 3rd edn., Cambridge: Cambridge University Press.
Smullyan, R. (1957) "Languages in which self reference is possible," *Journal of Symbolic Logic* 22, pp. 55–67.
Woodger, J. H. (trans.) (1956) *Logic Semantics, Metamathematics: Papers from 1923 to 1938 by Alfred Tarski*, Oxford: Clarendon Press.