Tarski

Benedict Eastaugh

April 30, 2015

1 Introduction

It is hard to overstate Alfred Tarski's impact on logic. Such were the importance and breadth of his results and so influential was the school of logicians he trained that the entire landscape of the field would be radically different without him. In the following chapter we shall focus on three topics: Tarski's work on formal theories of semantic concepts, particularly his definition of truth; set theory and the Banach–Tarski paradox; and finally the study of decidable and undecidable theories, determining which classes of mathematical problems can be solved by a computer and which cannot.

Tarski was born Alfred Tajtelbaum in Warsaw in 1901, to a Jewish couple, Ignacy Tajtelbaum and Rosa Prussak. During his university education, from 1918 to 1924, logic in Poland was flourishing, and Tarski took courses with many famous members of the Lvov–Warsaw school, such as Tadeusz Kotarbiński, Stanisław Leśniewski and Jan Łukasiewicz. Prejudice against Jews was widespread in interwar Poland, and fearing that he would not get a faculty position, the young Alfred Tajtelbaum changed his name to Tarski. An invented name with no history behind it, Alfred hoped it would sound suitably Polish. The papers confirming the change came through just before completing his doctorate (he was the youngest ever to be awarded one by the University of Warsaw), and he was therefore awarded the degree under his new name of Alfred Tarski.

Struggling to obtain a position in line with his obvious brilliance, Tarski took a series of poorly-paid teaching and research jobs at his alma mater, supporting himself by teaching high-school mathematics. It was there that he met the woman who would become his wife, fellow teacher Maria Witkowska. They married in 1929, and had two children: their son Jan was born in 1934, and their daughter Ina followed in 1938. Passed over for a professorship at the University of Lvov in 1930, and another in Poznan in 1937, Tarski was unable to secure the stable employment he craved in Poland.

Despite these professional setbacks, Tarski produced a brilliant series of publications throughout the 1920s and 1930s. His work on the theory of truth laid the ground not only for model theory and a proper understanding of the classical logical consequence relation, but also for research on the concept of truth that is still bearing fruit today. Tarski's decision procedure for elementary algebra and geometry, which he regarded as one of his two most important contributions, was also developed in this period.

In 1939 he took ship to the United States for a lecture tour, with a thought of finding employment there. Seemingly oblivious to the impending conflagration, Tarski nevertheless contrived to escape mere weeks before war with Germany broke out, but leaving his wife and children behind. Working as an itinerant lecturer at Harvard, the City College of New York, Princeton and Berkeley, Tarski spent the war years separated from his family.

Back in Poland, Maria, Jan and Ina were taken into hiding by friends. Despite intermittent reports that they were still alive, Tarski spent long periods without news, and his attempts to extricate them from Poland were all in vain. It was not until the conclusion of the war that he learned that while his wife and children had survived, most of the rest of his family had not. His parents perished in Auschwitz, while his brother Wacław was killed in the Warsaw Uprising of 1944. About thirty of Tarski's close relatives were amongst the more than three million Polish Jews murdered in the Holocaust, along with many of his colleagues and students, including the logician Adolf Lindenbaum and his wife, the philosopher of science Janina Hosiasson-Lindenbaum.

In 1945, Tarski gained the permanent position he craved at UC Berkeley, where Maria and the children joined him in 1946. Made professor in 1948, Tarski remained in California until his death in 1983. There he built a school in logic and the philosophy of science and mathematics that endures to this day: a testament to his brilliance as a scholar, his inspirational qualities as a teacher, and his sheer force of personality.

The most universally known and acclaimed part of Tarski's career consists of his work on the theory of truth, so it is natural that we begin our journey there. Section 2 starts from the liar paradox, and then turns to Tarski's celebrated definition of truth for formalised languages. This leads us to the undefinability theorem: that no sufficiently expressive formal system can define its own truth predicate.

Much of Tarski's early research was in set theory. Although he remained interested in the area for the rest of his working life, his best-known contribution to the field remains the paradoxical decomposition of the sphere which he developed in collaboration with Stefan Banach, colloquially known as the Banach–Tarski paradox. This striking demonstration of the consequences of the Axiom of Choice is explored in section 3.

As a logician, only Kurt Gödel outshines Tarski in the twentieth century. His incompleteness theorems are the singular achievement around which the story of section 4 pivots. Before Gödel, logicians still held out hope for a general algorithm to decide mathematical problems. Many of this area's successes in the 1920s are due to Tarski and his Warsaw students, such as the discovery that when formulated in a language without the multiplication symbol, the theory of arithmetic is decidable. In 1936, five years after Gödel's discovery of incompleteness, Alonzo Church and Alan Turing showed that the general decision problem for first-order logic was unsolvable. The focus then turned from complete, decidable systems to incomplete, undecidable ones, and once again Tarski and his school were at the forefront. Peano arithmetic was incomplete and undecidable; how much could it be weakened and retain these properties? What were the lower bounds for undecidability?

This is only intended as a brief introduction to Tarski's life and work, and as such there are many fascinating results, connections and even whole areas of study which must go unaddressed. Fortunately the history of logic has benefitted in recent years from some wonderful scholarship. The encyclopaedic *Handbook of the History of Logic* is one such endeavour, and Keith Simmons's chapter on Tarski [Simmons 2009] contains over a hundred pages. Tarski is also the subject of an engrossing biography by Anita Burdman Feferman, together with her husband and Tarski's former student, Solomon Feferman [Feferman and Feferman 2004]. Entitled *Alfred Tarski: Life and Logic*, it mixes a traditional biography of Tarski's colourful life with technical interludes explaining some of the highlights of Tarski's work. Finally, in addition to being a logician of the first rank, Tarski was an admirably clear communicator. His books and papers, far from being of merely historical interest, remain stimulating reading for logicians and philosophers. Many of them, including early papers originally published in Polish or German, are collected in the volume *Logic, Semantics, Metamathematics* [Tarski 1983].

2 The theory of truth

Semantic concepts are those which concern the meanings of linguistic expressions, or parts thereof. Amongst the most important of these concepts are *truth*, *logical consequence* and *definability*. All of these concepts were known in Tarski's day to lead to paradoxes. The most famous of these is the *liar paradox*. Consider the sentence "Snow is white". Is it true, or false? Snow is white: so the sentence "Snow is white" is true. If snow were not white then it would be false. Now consider the sentence "This sentence is false". Is it true, or false? If it's true, then the sentence is false. But if it's false, then the sentence is true. So we have a contradiction whichever truth value we assign to the sentence. This sentence is known as the *liar sentence*.

In everyday speech and writing, we appear to use truth in a widespread and coherent way. Truth is a foundational semantic concept, and therefore one which we might naively expect to obtain a satisfactory philosophical understanding of. The liar paradox casts doubt on this possibility: it does not seem to require complex or far-fetched assumptions about language in order to manifest itself, but instead arises from commonplace linguistic devices and usage such as our ability to both use and mention parts of speech, the property of bivalence and the typical properties we ascribe to the truth predicate such as disquotation.

2.1 Tarski's definition of truth

Both their apparent ambiguity and paradoxes like the liar made mathematicians wary of semantic concepts. Tarski's analyses of truth, logical consequence and definability for formal languages thus formed major contributions to both logic and philosophy. This paved the way for model theory and much of modern mathematical logic on the one hand; and renewed philosophical interest in these semantic notions—which continues to this day—on the other.

In his seminal 1933 paper 'On the Concept of Truth in Formalized Languages' [1933], Tarski offered an analysis of the liar paradox. To understand Tarski's analysis, we first need to make a few conceptual points. The first turns on the distinction between use and mention. If we were to say that Tarski was a logician, we would be *using* the name "Tarski"—but if we said that "Tarski" was the name that logician chose for himself, we would be *mentioning* it. In the written forms of natural language we often distinguish between using a term and mentioning it by quotation marks. When we say that "Snow is white" is true if, and only if, snow is white, we both use and mention the sentence "Snow is white".

The liar paradox seems to rely on our ability not merely to use sentences—that is, to assert or deny them—but on our ability to refer to them. The locution "This sentence" in the liar sentence refers to (that is, mentions) the sentence itself, although it does not use quotation marks to do so. Consider the following variation on the liar paradox, with two sentences named A and B. Sentence A reads "Sentence B is false" while sentence B reads "Sentence A is true". We reason by cases: either sentence A is true, or it is false. If A is true, then B is false, so it is false that A is true—hence A is false, contradicting our assumption. So A must be false. But if A is false, then it is false that B is false, and so B is in fact true. B says that A is true, contradicting our assumption that it is true. So we have a contradiction either way.

In his analysis of the liar paradox, Tarski singles out two key properties which a language must satisfy in order for the paradox to occur in that language. The first consists of three conditions: the language must contain names for its own sentences; it must contain a semantic predicate "x is true"; and all the sentences that determine the adequate usage of the truth predicate must be able to be stated in the language. These conditions are jointly known as *semantic universality*.

The second property is that the ordinary laws of classical logic apply: every classically valid inference must be valid in that language. Tarski felt that rejecting the ordinary laws of logic

would have consequences too drastic to even consider this option, although many philosophers since have entertained the possibility of logical revision; see section 4.1 of Beall and Glanzberg [2014] for an introductory survey. Since a satisfactory analysis of truth cannot be carried out for a language in which the liar paradox occurs—as it is inconsistent—Tarski concluded that we should seek a definition of truth for languages that are not semantically universal.

There are different ways for a language to fail to be semantically universal. Firstly, it could fail to have the expressive resources necessary to make assertions about its own syntax: it could have no names for its own expressions. Secondly, it could fail to contain a truth predicate. Finally, the language might have syntactic restrictions which restrict its ability to express some sentences determining the adequate usage of the truth predicate.

This seems to exclude the possibility of giving a definition of truth for natural languages. Not only are they semantically universal—quotation marks, for instance, allow us to name every sentence of English within the language—but they actually *aim* for universality. If a natural language fails to be semantically universal then it will be expanded with new semantic resources until it regains universality. Tarski goes so far as to say that "it would not be within the spirit of [a natural language] if in some other language a word occurred which could not be translated into it" [Tarski 1983, p. 164]. When English fails to have an appropriate term to translate a foreign one, in cases like the German "schadenfreude" or the French "faux pas", the foreign term is simply borrowed and becomes a loanword in English.

Tarski therefore offered his definition of truth only for *formal languages*. These tend to be simpler than natural languages, and thus they are more amenable to metalinguistic investigation. The particular example that Tarski used was the calculus of classes, but essentially the same approach can be used to define truth for any formal language. As is standard in the current literature on formal theories of truth, we shall use the language of arithmetic.

A formal language is typically constructed by stipulating two main components. The first is the *alphabet*: the collection of symbols from which all expressions in the language are drawn. In the case of a first-order language like that of arithmetic, the alphabet includes (countably infinitely many) variables $v_0, v_1, \ldots$; logical constants $\forall, \exists, \neg, \wedge, \vee, \rightarrow, \leftrightarrow, =$; and punctuation (,). This is then enriched by the addition of non-logical constants, function symbols and relational predicates. In the case of the first-order language of arithmetic this includes the constant symbols 0 and 1; the two binary function symbols $+$ and $\times$; and the binary relation symbol $<$. The second component of a formal language is the *formation rules*, which state how one may build up well-formed formulas from the symbols of the alphabet. Again, in the case of arithmetic, these are just the standard recursive definitions familiar from first-order logic.[1]

A formal language is generally understood as one which can be expressed in terms of an alphabet and a set of formation rules. These rules are decidable: given a sequence s of symbols drawn from the alphabet, one can always determine in a finite number of steps whether or not the sequence is a well-formed formula of the language or not.

Implicit in the circumscription of a formal system—its syntax, its semantics, its axioms and rules of inference—is the idea of the metatheory in which all of these things are laid down. One of the major innovations in logic during the first part of the twentieth century was the recognition of this fact, and the subsequent results obtained by formalising the metatheory. Tarski was one of the pioneers in this area.

We call the language for which a definition of truth is to be given the *object language*, and the language in which we do so the *metalanguage*. The metalanguage can simply be an expansion of the object language with the necessary semantic terms, although this is not essential, as long as it contains translations of the terms of the object language. The metatheory is a theory—and

[1] See for example section 2.1 of Enderton [2001, pp. 69–79].

as Tarski showed, it can be a formal theory, i.e. a set of sentences in a formal metalanguage—in which to theorise about the object theory.

Here we pause to highlight a change in terminology between Tarski's work and current usage. Tarski uses the term "language" to denote what we nowadays call a *formal system*: not just a formal language in the sense described above, but also a set of axioms and inference rules associated with that language. In keeping with current practice, and fixing the logic throughout to be first-order classical logic, we shall be concerned with *theories*: sets of sentences of a given formal language, such as Peano arithmetic or ZFC. When Tarski writes of the "object language" and the "metalanguage" he thereby means what we mean when we write *object theory* and *metatheory*.

As we have seen, quotation marks are not the only linguistic device by which we can refer to sentences and their components. Demonstratives and names can both be used, but in his account of truth Tarski settled on *structural-descriptive names*: names for primitive parts of language which can be combined to yield the names of compound expressions. In the particular case of arithmetic, this amounts to providing a formal counterpart of the description of the language of arithmetic given above. It must contain names for variables $x_1, x_2, \ldots$; names for logical vocabulary such as $\neg, \wedge, \forall$; and names for grammatical symbols such as (and). It must also provide names for the nonlogical symbols: $0, 1, +, \times, <$. With the referential devices in hand, it must also provide a way to combine them to give the names of complex expressions such as *terms* (denoting expressions such as $1 \times (1 + 1)$), *atomic formulas* like $1 = 1$, and complex formulas like $1 = 0 \rightarrow 1 < 0$.

As Corcoran notes in his introduction to [Tarski 1983], Tarski effectively provided the first formal theory of syntax, something which has gone on to become a subject of substantial importance in computer science. Tarski's approach of adding a syntax theory to the object language is currently undergoing a small revival, being used in recent work by Leigh and Nicolai [2013]. However, we shall not pursue this method further, since in most applications within logic it has been superseded by an alternative which is available in arithmetic and other suitably expressive formal systems: *Gödel coding*.

So called because it was invented by Kurt Gödel in the course of his proof of the incompleteness theorems, Gödel coding is a way of encoding sentences in the language of arithmetic as particular natural numbers, in such a way that given any number coding a sentence we can determine just what that sentence is. The details of Gödel coding can be found in the previous chapter on Kurt Gödel; for our purposes all we need to know is that for any sentence φ in the language of arithmetic, its Gödel code $\ulcorner \varphi \urcorner$ is a natural number, denoted by a closed term of the language—that is to say by a numeral $\overline{n}$.

Tarski stressed two qualities which any definition of truth must satisfy: *formal correctness* and *material adequacy*. The former concerns the form of the definition, namely whether it provides an explicit definition of the predicate "is true"; the latter, whether the formal definition captures our informal concept of truth.

Think back to our example: "Snow is white" is true if snow is white, and false if snow is not white. In other words we have an equivalence: "Snow is white" is true if, and only if, snow is white. More generally, let S be a sentence and s a name for S. Then s is true if, and only if, S. This is Tarski's T-schema. Using Gödel coding, we can express this scheme in the formal language of arithmetic (plus the truth predicate):

$$\text{(T)} \qquad T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi.$$

The material adequacy condition Tarski argued for is called *Convention T*. According to Convention T, a materially adequate theory of truth for a language $\mathcal{L}$ should entail every sentence of

the T-schema for that language, and every Gödel code falling under the extension of the truth predicate T should stand for a sentence.

Having determined the properties that a successful definition of truth should satisfy, Tarski proceeded to present his definition. The first thing to note is that Tarski defines truth in terms of another semantic notion: *satisfaction*. For a full formal definition the reader should consult an introductory logic textbook.[2] The crucial idea is that satisfaction is a generalisation of truth, from sentences to all well-formed formulas of the language, including those with free variables. A satisfaction relation obtains between three components: a model $\mathfrak{M}$; an assignment s of elements of the domain of $\mathfrak{M}$ to free variables of the language $\mathcal{L}_{\mathfrak{M}}$ of $\mathfrak{M}$; and a formula φ in the language $\mathcal{L}_{\mathfrak{M}}$. In symbols we write this as

$$\mathfrak{M} \models \varphi[s].$$

Satisfaction is defined *recursively*, so for example a model $\mathfrak{M}$ and an assignment s satisfy a conjunction $\varphi \wedge \psi$ if, and only if, φ is satisfied by $\mathfrak{M}$ and s, *and* ψ is satisfied by $\mathfrak{M}$ and s. Consider the following example, where we fix the model to be the standard natural numbers $\mathbb{N}$. Take the formula $\varphi = v_1 < 1 \wedge v_2 = 0$, and a satisfaction function s_1 such that $s_1(v_1) = 0$ and $s_1(v_2) = 0$. Then we can see that the left conjunct "$v_1 < 1$" is satisfied by s_1 (and $\mathbb{N}$), since $s_1(v_1) < 1$, and so is the right conjunct "$v_2 = 0$", since $s_1(v_2) = 0$. Therefore the entire formula φ is satisfied by s_1.

Truth is defined as the limit case where no free variables appear in a formula: given some model $\mathfrak{M}$, a sentence φ is true in $\mathfrak{M}$ if, and only if, for every assignment s, $\mathfrak{M}$ and s satisfy φ. In the specific case of the language of arithmetic and the standard natural numbers $\mathbb{N}$, a sentence ψ in the language of arithmetic is true in $\mathbb{N}$ if, and only if, every assignment s of natural numbers to variables satisfies ψ. In symbols we can write this as

$$T(\ulcorner \varphi \urcorner) \Leftrightarrow (\forall \text{ assignments } s)\ \mathbb{N} \models \varphi[s].$$

Tarski constructed definitions of truth, in terms of satisfaction, for several different formal systems. As he noted in his 1933 paper, the method is entirely general. Using it we can define truth for arbitrary models and languages, and indeed this is one of the building blocks of mathematical logic as it stands today—in no small part due to Tarski's contributions.

2.2 Tarski's undefinability theorem

Gödel's incompleteness theorems (see the previous chapter) showed that no sufficiently strong, recursively axiomatizable theory of arithmetic S is *complete*, in the sense that there are sentences φ in the language of arithmetic such that neither φ nor its negation $\neg\varphi$ can be proved from the axioms of S. At the heart of this result is a sentence in the language of arithmetic, known as the *Gödel sentence*, which is a close relative of the liar sentence. Rather than asserting its own falsity, like the liar sentence, the Gödel sentence asserts its own *unprovability*.

Tarski's undefinability theorem shows something stronger: not only are consistent formal theories of a certain strength incomplete, but they cannot define the truth predicate for the language in which they are written. In other words, they cannot prove all instances of the T-schema for that language. The requirement that such theories be consistent is important: classical logic has a property called *explosion*, which means that if a theory S is inconsistent then it proves every sentence in the language of S, including every instance of the T-schema.

The following way of stating of Tarski's theorem is quite standard. The theory Q mentioned in the statement of the theorem is a very weak theory in the language of arithmetic. It was

[2] For example pages 80 to 86 of Enderton [2001].

discovered by Tarski's colleague, Raphael Robinson, and we shall learn more about it in section 4.3.

Undefinability Theorem. Let $\mathcal{L}$ be a language that extends the language of arithmetic $\mathcal{L}_{\mathrm{PA}}$. Suppose S is a consistent, recursively enumerable theory in $\mathcal{L}$ that includes the axioms of Robinson's Q. Given a Gödel numbering of the sentences of $\mathcal{L}$, there is no predicate τ definable in the language of arithmetic such that S proves the following equivalence scheme for all sentences φ of $\mathcal{L}$:

$$\tau(\ulcorner\varphi\urcorner) \leftrightarrow \varphi.$$

This might seem a little contradictory: the previous section spelled out in detail Tarski's definition of truth, but the undefinability theorem shows that truth cannot be defined. The resolution of this apparent conflict may already be evident, lying as it does in the way Tarski resolves the liar paradox: by stipulating that the definition of truth for a language $\mathcal{L}$ is not made in the object theory—this is ruled out by the undefinability theorem—but in the metatheory. To take our standard example, arithmetical truth is not definable in first-order arithmetic, but it is definable in a stronger theory, such as that of set theory or second order arithmetic.

Tarski emphasises in the historical notes at the end of Tarski [1933] that his work on truth was done independently and was largely complete, including the definition of truth, by 1929. After Gödel published his incompleteness theorems [Gödel 1931] Tarski realised that Gödel's methods could be used to prove the undefinability theorem.

3 The Banach–Tarski paradox

Tarski's work on the theory of truth was not the first time he had flirted with paradox. In 1924 he published a paper with fellow Polish mathematician Stefan Banach showing that a sphere could be cut up into finitely many pieces, and that those pieces could then be reassembled—using only translations and rotations—into two spheres, each with exactly the same volume as the original sphere.

This is, to say the least, a counterintuitive result. We expect Euclidean geometry to respect our basic physical intuitions (suitably idealised). If we take a knife and cut up an orange, we can't reassemble it into two oranges with the same size as the original one. The problem is *volume*, which in the case of the orange remains invariant no matter how we cut it up. Bounded sets, such as spheres, are supposed to have fixed, finite volumes: we can't just get a little extra from somewhere.

But the Banach–Tarski theorem shows that we can, after all, do exactly that, as long as we can cut up our sphere into parts which do not have well-defined volumes. These strange objects are called non-measurable sets, since they lack a *measure*: intuitively, a way of assigning a size to a bounded set. In the case of a line, this is just the length of an interval, like the set of all points between 0 and 1. For a plane it's the area which a bounded set encompasses; and in the three-dimensional case, its volume. While there are non-measurable subsets of $\mathbb{R}$ and $\mathbb{R}^2$, the one- and two-dimensional versions of the Banach–Tarski theorem are false, although a weaker version where the bounded set is cut into countably infinitely (rather than finitely) many pieces is true.

Non-measurable sets had been around since well before Tarski; the first proof of their existence was given by Giuseppe Vitali in 1905. Both Vitali's proof and the Banach–Tarski theorem rely on a set theoretic axiom whose use had been controversial ever since its introduction by the German set theorist Ernst Zermelo: the Axiom of Choice, or AC.

Definition. The *Axiom of Choice* is the statement that for every nonempty family of sets $\mathcal{F}$, there is a function f such that $f(S) \in S$ for every $S \in \mathcal{F}$.

Such an f is called a *choice function*, because it "chooses" an element from every set in the family. For finite families of sets, we can deduce the existence of choice functions from the other axioms of set theory. But once infinite collections are brought into the picture, the Axiom of Choice must be added as an additional postulate to guarantee the existence of choice functions for every nonempty family of sets.

The mathematicians of the day had two main quarrels with the Axiom of Choice. The first was that it allows one to prove a number of puzzling and deeply counterintuitive theorems, of which the Banach–Tarski theorem is the quintessential example. The second addressed the character of the axiom itself. In the presence of Zermelo's other axioms, it allows one to prove the existence of a great many sets, yet gives no way to define them. It is in this sense that the Axiom of Choice is referred to as a nonconstructive axiom.

The nature of choice functions therefore remains somewhat mysterious. A classic example can be found by comparing the real numbers $\mathbb{R}$ with the natural numbers $\mathbb{N}$. We first introduce the technical notion of a *wellorder*. A set X is wellordered by an ordering $\prec$ if and only if every nonempty subset $Y \subseteq X$ has a least element: some $x \in Y$ such that every $y \in Y$ is greater than or equal to x under the ordering $\prec$. It's easy to see that the natural numbers are wellordered under their natural ordering:

$$0 < 1 < 2 < \ldots < n < n + 1 < \ldots$$

However, the usual ordering on $\mathbb{R}$ is not a wellordering. To see this, consider the open interval $(0, 1) \subseteq \mathbb{R}$, which consists of all real numbers greater than 0 but less than 1. If the reals were wellordered then there would be a smallest real $x \in (0, 1)$. But $\frac{x}{2}$ is also a real number greater than 0 but less than 1, and $\frac{x}{2} < x$. So x could not be the smallest element of $(0, 1)$ after all.

So much for the usual ordering. But perhaps there is another way we can order the continuum to get a wellorder? After all, the rational numbers under their usual ordering are not wellordered—but there is another ordering on them which is a wellordering. As it turns out, the answer to this is no: the axioms of ZF alone do not imply that the continuum is wellordered. To prove that it is, we require the Axiom of Choice. In the theory obtained by adding AC to the axioms of ZF, known as ZFC, we can prove that there is a relation $\prec$ on $\mathbb{R}$ which wellorders the continuum. In fact, we can prove a lot more than that: ZFC proves the *Wellordering Principle*, which states that every set can be wellordered. And the relationship between AC and the Wellordering Principle doesn't stop there: if we assume only the axioms of ZF, plus the Wellordering Principle, we can prove the Axiom of Choice; they are equivalent.

But the wellordering principle merely says that there exists a wellordering; it doesn't tell us what the ordering is. In other words, it doesn't define it. Even worse, there may not even be such a definition: it's consistent with the axioms of ZFC that there is no formula in the language of set theory which defines a wellordering of the continuum, even though ZFC proves that such a wellordering exists.

Later developments in logic have given us a clearer view of what Banach and Tarski accomplished. Using Cohen's method of forcing, Robert M. Solovay constructed a model of the axioms of ZF plus the assertion that every subset of the real numbers $\mathbb{R}$ is measurable. In Solovay's model the Banach–Tarski theorem is false, showing that the Axiom of Choice is indeed required in order to prove it [Solovay 1970]. Raphael Robinson, Tarski's colleague at Berkeley, improved the Banach–Tarski theorem itself by showing that a paradoxical decomposition of the sphere could be achieved by cutting it into just five pieces—and that this is the minimum number possible [Robinson 1947]. More recently, Pawlikowski [1991] used work of Foreman and Wehrung

[1991] to show that the full strength of the Axiom of Choice is not required in order to prove the Banach–Tarski theorem: it suffices to assume a weaker principle, important in functional analysis, known as the Hahn–Banach theorem.

Giving a complete proof of the Banach–Tarski theorem is, unfortunately, outside the scope of this chapter. The reader interested in a fuller account should consult Jech [1973] for a relatively comprehensive reference. A recent popular account is Wapner [2005].

4 Decidable and undecidable theories

4.1 Mechanical mathematics and the Entscheidungsproblem

The history of the *decision problem*, or *Entscheidungsproblem*—the German name by which it is often known—can be traced back to Leibniz, whose hope it was to devise a mechanical means for deriving the truth or falsity of mathematical statements. The problem lay fallow until 1900, when in an address to the International Congress of Mathematicians, David Hilbert laid down a series of challenges to the mathematical community; these became known as *Hilbert's problems*. The tenth of these problems concerned the solubility of Diophantine equations:

> Given a diophantine equation with any number of unknown quantities and with rational integral numerical coefficients: *To devise a process according to which it can be determined by a finite number of operations whether the equation is solvable in rational integers.*[3]

Hilbert's tenth problem was not resolved until 1970, when Yuri Matiyasevich put in place the final pieces of a proof begun many years earlier by Julia Robinson (another of Tarski's students), Martin Davis and Hilary Putnam, and which showed that no such finite procedure exists. Developing the theme of Hilbert's tenth problem in a more general and precise way, Hilbert and Wilhelm Ackermann posed the classical version of the Entscheidungsproblem in [Hilbert and Ackermann 1928]. The solution to the decision problem would be an algorithm that, given a formal language $\mathcal{L}$ and a theory T written in that language, decided whether or not any particular sentence φ in the language $\mathcal{L}$ was true in all models of T. In other words, the algorithm should determine whether or not φ is a logical consequence of T.

As it stood, almost every part of the Entscheidungsproblem could be stated in formal terms: there were unambiguous mathematical definitions of the notions of a formal language and of a theory formulated in that language. Once Gödel proved his completeness theorem for first-order logic, it was also clear that for a sentence to be a consequence of a particular formal theory was precisely for it to be derivable from that theory in an appropriate formal calculus. However, the concept of an effective procedure or algorithm remained unformalised.

If, in the early 1930s, someone had come along with an algorithm solving the Entscheidungsproblem then it would have been clear to the mathematical community that it was in fact such an algorithm. But since they did not—and with Gödel's incompleteness theorems fresh in their minds—logicians turned their efforts towards proving that there could be no such algorithm. To do this, the fuzzy notion of an effective procedure needed to be given a precise formal definition. Otherwise any purported proof of the impossibility of solving the Entscheidungsproblem would have been vulnerable to the accusation that the proof did not cover all of the cases it needed to.

Various definitions were offered, from the notion of a *recursive function* developed by Jacques Herbrand and Kurt Gödel, to Alonzo Church's property of *λ-definability*. Church proved in 1936

[3] Hilbert [1902, p. 458], italics in original.

that the Entscheidungsproblem has no solution, if the notion of an effective procedure is identified with recursiveness. This identification, known as Church's Thesis, met with resistance in the logical community, not least from Gödel. In the end it was Alan Turing's conceptual analysis of computation, which led him to develop the idea of the Turing machine, that convinced Gödel to accept what is now known as the Church–Turing thesis: the functions that can be effectively computed are precisely those which can be computed by a Turing machine. As Turing and Stephen Kleene proved, this set of functions is identical to that picked out by the other formal notions of computability: recursiveness, λ-definability and Turing computability all coincide. The scientific consensus since then has sided with Gödel: effective computability is Turing computability, and Hilbert's Entscheidungsproblem is unsolvable.

4.2 Decidable theories

Well before Gödel's incompleteness theorems burst into the startled minds of the logical community in 1931, and almost a full decade before Church and Turing's negative resolution of the decision problem, Tarski had done pioneering work on *decidable* theories: ones where there is an algorithm that determines whether or not a given statement is a consequence of the theory. Decidability is closely linked to *completeness*, and proofs of one are often also proofs of the other.

Definition (completeness and decidability). A theory T in a language $\mathcal{L}$ is *complete* iff for every $\mathcal{L}$-sentence φ, either $T \vdash \varphi$ or $T \vdash \neg\varphi$. T is *decidable* iff there is a decision procedure that determines whether or not an $\mathcal{L}$-sentence φ is a theorem of T.

Many major advances were made during a research seminar at the University of Warsaw which Tarski ran from 1927 to 1929. The topic of the seminar was *quantifier elimination.* This is a technique in the field we now call model theory. It first emerged in Löwenheim [1915]'s work, and appeared in its full form in Skolem [1919].

Definition (quantifier elimination). A first-order theory T in a language $\mathcal{L}$ admits of *quantifier elimination* if for every $\mathcal{L}$-formula $\varphi(x_1, \ldots, x_n)$ there is a quantifier-free $\mathcal{L}$-formula $\varphi^*(x_1, \ldots, x_n)$ such that

$$T \vdash \varphi(x_1, \ldots, x_n) \leftrightarrow \varphi^*(x_1, \ldots, x_n).$$

Proving that a theory admits of quantifier elimination usually involves specifying an algorithm by which a formula φ containing quantifiers can be transformed into an equivalent formula φ^* without them. Quantifier-free formulas are built up by boolean combinations from atomic formulas, so if a theory proves or refutes every atomic sentence, then it proves or refutes every quantifier-free sentence too. Typically it is easier to show decidability for atomic sentences, since often this is simply a matter of computation, and if a theory with this property admits of quantifier elimination then a decidability result for the theory follows easily.

Langford [1927a,b] used this technique to solve the decision problem for the first-order theory of dense linear orders. In the seminar on quantifier elimination, Tarski began by extending Langford and Skolem's results, before turning to more ambitious targets. One of these was the decidability of the additive theory of the natural numbers. This theory is formulated in the language consisting of the constant symbols 0 and 1, and the binary function symbol $+$. Its axioms are the universal closures (that is, every free variable is bound by an outer universal

quantifier) of the following formulas:

(P1) $0 \neq n + 1$

(P2) $n + 1 = m + 1 \to n = m$

(P3) $n + 0 = n$

(P4) $n + (m + 1) = (n + m) + 1$

(P5) $(\varphi(0) \wedge \forall n(\varphi(n) \to \varphi(n + 1))) \to \forall n\varphi(n).$

Note that the final axiom is actually a scheme, where $\varphi(n)$ is any formula of this language containing one or more free variables. Tarski's student Mojżesz Presburger proved in 1928 that this theory—now known as Presburger arithmetic, in his honour—is consistent, complete, and decidable.[4] These results formed Presburger's Master's thesis at the University of Warsaw, and were his only published results in logic: he left the academy soon after, and like so many other Polish Jews, perished in the Holocaust [Zygmunt 1991].

The centrepiece of Tarski's research on this topic was his proof that the theory of real closed fields admits of quantifier elimination. Van den Dries [1988, p. 7] goes so far as to say that "Tarski made a fundamental contribution to our understanding of $\mathbb{R}$, perhaps mathematics' most basic structure." The theory of real closed fields is formulated in the *language of ordered rings*, $\mathcal{L}_{\text{or}}$, which is the language of rings (the constant symbols 0 and 1, and the binary function symbols $+$ and $\cdot$) supplemented with the order relation $\leq$. The *axioms for ordered fields* are the usual definitions of addition and multiplication for fields, with additional axioms governing the order relation:

$$\begin{aligned} a < b \vee a = b \vee b < a, \\ a \leq b \to a + c \leq b + c, \\ 0 \leq a \wedge 0 \leq b \to 0 \leq a \cdot b. \end{aligned}$$

A *real closed field* is an ordered field which also obeys the following continuity scheme: for every polynomial term p, if there exist real numbers a and b such that $a < b$ and $p(a) < 0 < p(b)$, then there exists another real number c such that $a < c < b$ and $p(c) = 0$. Tarski proved that given any formula $\varphi(x_1, \ldots, x_m)$ in the language of ordered rings, we can effectively find a quantifier-free formula $\varphi^*(x_1, \ldots, x_m)$ and a proof of the equivalence $\varphi \leftrightarrow \varphi^*$ that uses only the axioms for real closed fields.

Many important and fruitful consequences follow from this result. To begin with, the theory of real closed fields is both complete and decidable. Moreover, since the real numbers $\mathbb{R}$ are a real closed field—indeed, the prototypical one—it follows that the theory of $\mathbb{R}$ is also complete and decidable. More formally, given any sentence φ in the language $\mathcal{L}_{\text{or}}$ of ordered rings, there is a finite procedure that determines whether or not $\mathbb{R} \models \varphi$. This stands in striking contrast to the theory of the rational numbers $\mathbb{Q}$, which does not have this property. Julia Robinson proved that the integers $\mathbb{Z}$ are definable in the field $\mathbb{Q}$, and thus the theory of the rational numbers is neither complete nor decidable [Robinson 1949].

The solution of the decision problem for the theory of real closed fields was intimately linked to Tarski's work on geometry, and in particular to the axioms he gave for what he called *elementary geometry*: a substantial fragment of Euclidean geometry, formulated in first-order logic with identity, and requiring no set theory. By reducing the theory of elementary geometry to the

[4] The axioms above are closer to those in Hilbert and Bernays [1968, 1970] than the axioms used by Presburger himself, as noted by Zygmunt [1991, p. 221]. Presburger's original axioms can be found on pp. 218-9 of Zygmunt [1991].

theory of elementary algebra—that is to say, the theory of real closed fields—Tarski proved that it was decidable.

As with much of Tarski's work, his decision procedure for elementary algebra and geometry was worked out in the 1930s but publication was delayed until much later. An attempt to publish it in a French journal was ruined by the German invasion of 1940. It finally made it into print as a RAND Corporation report and was subsequently reprinted as [Tarski 1951]; an extensive discussion, with interesting historical as well as mathematical insights, is van den Dries [1988]. Tarski and his school proved many other decidability results, which have been surveyed by Doner and Hodges [1988]. A more general study of Tarski's work in model theory is Vaught [1986].

4.3 Undecidable theories

The success of Tarski and his students in classifying decidable theories notwithstanding, it was clear that undecidability was a widespread phenomenon. Presburger's theorem showed that one way of weakening Peano arithmetic, by removing multiplication, resulted in a decidable theory. A natural question to ask was thus: how weak could an undecidable subtheory of Peano arithmetic be?

A precise answer to this question was provided by Raphael Robinson, who formulated the weak theory of arithmetic Q—now known as *Robinson arithmetic*—and proved its undecidability. Robinson's Q is formulated in the language of first order arithmetic, that is, the language of first order logic supplemented by the constant symbol 0; the unary function symbol S denoting the successor function; and binary function symbols $+$ and $\cdot$ denoting addition and multiplication respectively.

Definition (Robinson's Q*).* The axioms of *Robinson arithmetic* or Q are the universal closures of the following.

(Q1) $$Sx \neq 0$$

(Q2) $$Sx = Sy \to x = y$$

(Q3) $$y = 0 \vee \exists x(Sx = y)$$

(Q4) $$x + 0 = x$$

(Q5) $$x + Sy = S(x + y)$$

(Q6) $$x \cdot 0 = 0$$

(Q7) $$x \cdot Sy = (x \cdot y) + x$$

Q has many interesting properties. For instance, it proves the commutativity of addition in every individual case: $t + s = s + t$ holds for all closed terms (those containing no free variables) t and s in the language of arithmetic. However, it cannot prove the universal generalisation $\forall x \forall y(x + y = y + x)$.

Definition (essential undecidability). A theory T is *decidable* if the set of its provable consequences in the language of T is recursive, and *undecidable* otherwise. A theory S is *essentially undecidable* if S is undecidable and every consistent extension of S is also undecidable.

Theorem (Robinson). Q is essentially undecidable.

This result was originally proved in 1939 for a stronger subsystem of first order Peano arithmetic by Tarski and Andrzej Mostowski. Raphael Robinson showed in [Robinson 1950] that it also holds for Q; for details see Tarski et al. [1953, pp. 39–40]. Robinson also proved the following intriguing theorem, showing that the essential undecidability of Q is in some sense *irreducible*.

Theorem (Robinson). None of the theories obtained by removing one of the 7 axioms of Q is essentially undecidable.

Tarski, Mostowski and Robinson collaborated on a book, *Undecidable Theories* [Tarski et al. 1953]. The slimness of this volume belies its importance: in it, Tarski not only set out a general and powerful method for proving undecidability results, but he inspired a new wave of research. The first part of the book consists of a general introduction to the issue of undecidability, written by Tarski. In it he uses the notion of an *interpretation* of one theory in another to develop a quite general method of proving undecidability results. This proceeds, as Tarski puts it, in an *indirect* manner: rather than directly demonstrating undecidability, as for example Church did, it proves that a theory T is undecidable (or essentially undecidable) because it interprets a theory S which is already known to be undecidable (or essentially undecidable).

Part II contains detailed proofs of a number of key undecidability results, including Robinson's theorems that Q is essentially undecidable, and was co-authored by Tarski, Mostowski and Robinson. One of the striking features of the results is the generality and clarity that Tarski achieved, analysing in great detail the results of the past twenty years and distilling them into a pure and powerful form. The clearest example of this is theorem 1 of this part, which states that no consistent theory T can define both the diagonal function (which should be familiar from the discussion in the preceding chapter of Gödel's incompleteness theorems) and the set of theorems of T.

In the final part, Tarski uses the machinery of interpretability developed in Part I, along with the undecidability results of Part II, to show that the first-order theory of groups is undecidable. Here once more we see the unity of Tarski's project, since in 1949 his student Wanda Szmielew had shown that the first-order theory of Abelian groups—formed by adding the commutativity axiom to the theory of groups—is *decidable*. The theory of groups is therefore not essentially undecidable, since it has a consistent decidable extension.

References

J. Beall and M. Glanzberg. Liar Paradox. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2014 edition, 2014.

J. Doner and W. Hodges. Alfred Tarski and Decidable Theories. *The Journal of Symbolic Logic*, 53(1):20–35, 1988.

H. B. Enderton. *A Mathematical Introduction to Logic*. Harcourt Academic Press, second edition, 2001.

A. B. Feferman and S. Feferman. *Alfred Tarski: Life and Logic*. Cambridge University Press, 2004.

S. Feferman, J. W. Dawson, Jr, S. C. Kleene, G. H. Moore, R. M. Solovay, and J. van Heijenoort, editors. *Kurt Gödel: Collected Works. I: Publications 1929–1936*. Oxford University Press, 1986.

M. Foreman and F. Wehrung. The Hahn–Banach theorem implies the existence of a non-Lebesgue measurable set. *Fundamenta Mathematicae*, 138:13–19, 1991.

K. Gödel. Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik Physik*, 38:173–198, 1931. English translation in van Heijenoort [1967, pp. 596–616] and in Feferman et al. [1986, pp. 144–195].

D. Hilbert. Mathematical problems. *Bulletin of the American Mathematical Society*, 8:437–479, 1902. doi: 10.1090/S0002-9904-1902-00923-3. Translated for the Bulletin by Mary Winston Newson.

D. Hilbert and W. Ackermann. *Grunzuge Der Theoretischen Logik*. Springer-Verlag, 1928.

D. Hilbert and P. Bernays. *Grundlagen der Mathematik*. Springer Verlag, 2nd edition, 1968, 1970. Volume I and Volume II.

T. J. Jech. *The Axiom of Choice*. Dover, 1973. Originally published by North–Holland, republished by Dover in 2008.

C. H. Langford. Some theorems on deducibility. *Annals of Mathematics, second series*, 28:16–40, 1927a.

C. H. Langford. Theorems on deducibility (second paper). *Annals of Mathematics, second series*, 28:459–471, 1927b.

G. E. Leigh and C. Nicolai. Axiomatic truth, syntax and metatheoretic reasoning. *The Review of Symbolic Logic*, 6(4):613–636, 2013. doi: 10.1017/S1755020313000233.

L. Löwenheim. Über Möglichkeiten im Relativkalkül. *Mathematische Annalen*, 76:447–470, 1915.

J. Pawlikowski. The Hahn-Banach theorem implies the Banach–Tarski paradox. *Fundamenta Mathematicae*, 138:21–22, 1991.

J. Robinson. Definability and decision problems in arithmetic. *The Journal of Symbolic Logic*, 14:98–114, 1949.

R. M. Robinson. On the decomposition of spheres. *Fundamenta Mathematicae*, 34:246–260, 1947.

R. M. Robinson. An essentially undecidable axiom system. In *Proceedings of the International Congress of Mathematicians*, volume 1, pages 729–730, 1950.

K. Simmons. Tarski's Logic. In D. M. Gabbay and J. Woods, editors, *Handbook of the History of Logic. Volume 5. Logic from Russell to Church*, pages 511–616. Elsevier, 2009.

T. Skolem. Untersuchungen über die Axiome des Klassenkalkuls und über Produktations- und Summationsprobleme, welche gewisse Klassen von Aussagen betreffen. *Skrifter utgit av Videnskapsselskapet i Kristiania. I, Matematisk-Naturvidenskabelig Klasse*, 3, 1919.

R. M. Solovay. A model of set-theory in which every set of reals is Lebesgue measurable. *Annals of Mathematics, Second Series*, 92(1):1–56, 1970.

A. Tarski. The Concept of Truth in Formalized Languages. In *Tarski 1983*, pages 152–278. 1933.

A. Tarski. *A Decision Method for Elementary Algebra and Geometry*. University of California Press, 1951. Prepared for publication by J. C. C. McKinsey. Orginally published in 1948 as RAND report R-109, RAND Corp., Santa Monica, CA.

A. Tarski. *Logic, Semantics, Metamathematics*. Hackett, 1983. Edited by J. Corcoran. 2nd revised edition. Original 1956 edition translated and edited by J. H. Woodger.

A. Tarski, A. Mostowski, and R. M. Robinson. *Undecidable Theories*. North-Holland, 1953.

L. van den Dries. Alfred Tarski's Elimination Theory for Real Closed Fields. *The Journal of Symbolic Logic*, 53(1):7–19, 1988.

J. van Heijenoort, editor. *From Frege to Gödel: A Source Book in Mathematical Logic, 1879–1931*. Harvard University Press, 1967.

R. L. Vaught. Alfred Tarski's Work in Model Theory. *The Journal of Symbolic Logic*, 51(4): 869–882, 1986. doi: 10.2307/2273900.

L. M. Wapner. *The Pea and the Sun: A Mathematical Paradox*. A. K. Peters, 2005.

J. Zygmunt. Mojżesz Presburger: Life and Work. *History and Philosophy of Logic*, 12(2):211–223, 1991. doi: 10.1080/014453409108837186.