High-School Algebra and the Limits of AI*

Alexander V. Gheorghiu MIMA, University of Southampton

Tarski's puzzle

You are back in your high-school maths class and learning the laws of arithmetic: addition is commutative, multiplication distributes over addition, and so on. Your teacher lists the rules on the blackboard:

$$\begin{aligned} x+y=y+x, \qquad & (x+y)+z=x+(y+z),\\ x\cdot 1=x, \qquad & x\cdot y=y\cdot x,\\ (x\cdot y)\cdot z=x\cdot(y\cdot z), \qquad & x\cdot(y+z)=x\cdot y+x\cdot z,\\ 1^x=1, \qquad & x^1=x,\\ x^{y+z}=x^y\cdot x^z, \qquad & (x\cdot y)^z=x^z\cdot y^z,\\ (x^y)^z=x^{y\cdot z}. & \end{aligned}$$

While these rules were certainly known long before, they can be firmly traced back to Richard Dedekind [1] in the 19th century.

As a curious student you throw your hand up and ask, 'Miss, can every problem be solved using them?' (meaning, identities involving only addition, multiplication and exponentiation over the positive integers). This question was originally posed by Alfred Tarski in the 1960s. In the 1970s, Dana Scott mentioned the problem in passing to a young student named Alex Wilkie.

After becoming a professor of mathematics at Yale University, Wilkie showed that the answer is: no. He presented the following equation as a counter example:

$$\begin{aligned} \left((1+x)^y+(1+x+x^2)^y\right)^x\cdot\left((1+x^3)^x+(1+x^2+x^4)^x\right)^y \\ =\left((1+x)^x+(1+x+x^2)^x\right)^y \\ \cdot\left((1+x^3)^y+(1+x^2+x^4)^y\right)^x. \end{aligned}$$

This identity is valid for all positive integers, as can be seen by factoring $(1-x+x^2)^{xy}$ from both sides, but it cannot be derived using just the 11 high-school identities given above. Without subtraction, there is no way to express or manipulate the underlying structure of Wilkie's identity to prove it. See Burris and Lee [2] for a more complete account.

This shows that there is a 'gap' left by the rules given in your high-school class. That is, there is a true arithmetic statement that cannot be proved from them. Very well, so then what would be a complete set? Gurevič [3] showed that no finite list of identities could be complete. Of course, there are trivial examples of infinite lists that are complete. Take, for example the set of all true equations over the positive integers, but these defeat the purpose of having the list in the first place.

Hilbert's dream

Tarski's puzzle is a microcosm of a much grander problem known as *Hilbert's programme*. In the early 20th century, David Hilbert articulated an ambitious vision – to formalise all of mathematics within a small, consistent set of axioms and logical rules. He was thinking about *The Elements* by Euclid. With five 'postulates' and five 'common notions', Euclid gives a foundation to all of classical geometry. This is a major work and exemplifies the mathematical virtues of rigour, clarity and elegance. Hilbert thought that this should serve as a model of all of mathematics.

Let's return to arithmetic and simplify the problem by removing exponentiation. We want some list $\mathfrak{A}$ of laws that completely govern the ring $\langle\mathbb{N},+,\cdot\rangle$. These *axioms* $\mathfrak{A}$ should be 'sound' in the sense that they are true of $\mathbb{N}$ and 'complete' in the sense that any true statement ϕ about $\mathbb{N}$ (e.g., $(x+1)^2=x^2+2x+1$) can be proved using them. Is such a list possible?

As with Tarski's problem, no finite list will do. Nonetheless, we may hope for something *finitary* in the sense of *recursively enumerable*. That is, we have a surjection $f:\mathbb{N}\to\mathfrak{A}$ that can be defined by a program on a standard computer (e.g., using Python on your laptop) and list out the axioms as we need them. Hilbert's dream would be satisfied if we could find such a finitary $\mathfrak{A}$ and prove that it is complete for $\mathbb{N}$. This dream was famously upended by the Austrian-born mathematician Kurt Gödel.

Gödel's challenge

In 1924, Gödel enrolled to study physics at the University of Vienna. Attending lectures by the philosopher Rudolf Carnap, Gödel became increasingly interested in the foundations of mathematics. He observed that a very weak, finite list of axioms called Robinson's $\mathfrak{Q}$ has a curious feature: it could represent mathematical reasoning as arithmetic expressions. Assuming $\mathfrak{Q}$ was consistent and complete, this would yield a contradiction. Not a contradiction within $\mathfrak{Q}$ but within mathematics.

Theorem (Gödel's first incompleteness theorem)**.** *Any recursively enumerable list of axioms $\mathfrak{A}\supseteq\mathfrak{Q}$ is incomplete. That is, there is an arithmetic sentence g that $\mathfrak{A}$ neither proves nor disproves.*

This result has captured the public imagination, but it is often misrepresented. It does not mean that mathematics is broken or fundamentally limited nor does it imply any general limits on what is knowable. Rather, it shows that there are limits within any formal system capable of expressing basic arithmetic. Gödel's proof of this theorem is subtle, devious and elegant. It comes in three acts.

Act I: Representation.

The first act is the arithmetisation of provability. Suppose we have a finitary set of axioms $\mathfrak{A}$ containing Robinson's $\mathfrak{Q}$. We want to represent a $\mathfrak{A}$-proof $\mathcal{P}$ as an arithmetic expression $\ulcorner\mathcal{P}\urcorner$ (e.g., a number) using an encoding that can be done using the tools in $\mathfrak{A}$. Gödel realised he could do it using $\mathfrak{Q}$, which is by assumption contained in $\mathfrak{A}$.

To begin, observe that the language of arithmetic contains denumerably many symbols, namely, addition, multiplication, equality, parentheses, variables, constants and so on. To each of these symbols, we assign a unique natural number, which we call the 'code' of that symbol. For example, we might assign numbers 1, 2, 3 and 4 to the symbols 0, 1, + and =, respectively.

* Graham Hoare Prize 2025 winning article

An arithmetic statement ϕ is a finite sequence of such symbols together with logical terminology such as 'and' and 'for any'. For example, if $\phi :=$ '$0 + 1 = 1$', then it is captured by the sequence $(1, 3, 2, 4, 2)$. We can use $\mathfrak{Q}$ and the *fundamental theorem of arithmetic* (FTA) to encode such sequences into whole numbers. For example, ϕ is encoded as

$$\ulcorner \phi \urcorner := 2^1 \cdot 3^3 \cdot 5^2 \cdot 7^4 \cdot 11^2.$$

This number $\ulcorner \phi \urcorner$ is known as the *Gödel number* of ϕ. Importantly, given a whole number n, we can use $\mathfrak{Q}$ to decode it. That is, given a number n that is an encoding of a formula ϕ, we can recover ϕ from n.

If we can encode arithmetic sentences in this way, we can also encode proofs of them. Hilbert [4] and others (cf., Gentzen [5]) developed a theory of *proofs* as mathematical objects. In this theory, a proof $\mathcal{P}$ using the laws $\mathfrak{A}$ can be viewed as a finite sequence of arithmetic statements $(\phi_1, \phi_2, \ldots, \phi_n)$ in which each ϕ_i is justified by some ϕ_j for $j < i$ or by $\mathfrak{A}$. Using the FTA again, we have

$$\ulcorner P \urcorner := 2^{\ulcorner \phi_1 \urcorner} \cdot 3^{\ulcorner \phi_2 \urcorner} \cdot \ldots \cdot p_n^{\ulcorner \phi_n \urcorner},$$

where p_n is the nth prime number.

Let $\mathfrak{A} \vdash \phi$ denote that there is a $\mathfrak{A}$-proof of the arithmetic statement ϕ. Gödel constructed an arithmetic sentence $\beta(x, y)$ such that $\mathfrak{A} \vdash \beta(m, n)$ if and only if (iff) m and n are the Gödel numbers of a formula ϕ and a list of formulae $\mathcal{P}$ such that $\mathcal{P}$ is an $\mathfrak{A}$-proof of ϕ. (The β stands for '*Beweis*', the German word for 'proof'.) In this way, Gödel successfully encoded provability using $\mathfrak{A}$ within $\mathfrak{A}$ itself.

Act II: Diagonalisation.

The second act was to show that this encoding enables a kind of self-reference. Gödel was inspired by the *liar paradox*: the statement 'This sentence is false.' Gödel realised he could encode this kind of self-referential paradox within $\mathfrak{A}$ about $\mathfrak{A}$-provability. To do this, he introduced the diagonal lemma.

Lemma (Diagonal lemma)**.** *For any formula $\phi(x)$ in the language of arithmetic, there exists a formula δ such that $\mathfrak{A} \vdash \delta$ iff $\mathfrak{A} \vdash \phi(\ulcorner \delta \urcorner)$.*

The actual statement is a little stronger. While the proof is technical, it is only a paragraph long, as Gödel constructs δ parametrically on ϕ. It is essentially a version of Cantor's method of diagonalisation [6] used to show that the real numbers are uncountable. This sets the stage for the final act.

Act III: Incompleteness.

We express $\mathfrak{A}$-provability as $\exists x \beta(x, y)$, which says that for some number n we have $\beta(n, y)$. Such an n is the Gödel number of a proof in $\mathfrak{A}$ of the proposition whose Gödel number is y. Then we express $\mathfrak{A}$-*unprovability* $\gamma(y)$ as its negation, *not* $\exists x \beta(x, y)$. Intuitively, $\gamma(n)$ says that n is the Gödel number of a formula that can *not* be proved from $\mathfrak{A}$.

Applying the diagonal lemma to $\gamma(y)$ produces a formula g such that $\mathfrak{A} \vdash g$ iff $\mathfrak{A} \vdash \gamma(\ulcorner g \urcorner)$. This causes a contradiction:

- If g is true, then $\mathfrak{A} \vdash g$ since $\mathfrak{A}$ is complete. Therefore, there is a proof $\mathcal{P}$ of g from $\mathfrak{A}$ and $\mathfrak{A} \vdash \beta(\ulcorner P \urcorner, \ulcorner g \urcorner)$. But, by construction of g, we also have $\mathfrak{A} \vdash \gamma(\ulcorner g \urcorner)$. That is, $\mathfrak{A} \nvdash \beta(n, \ulcorner g \urcorner)$ for any n. This is a contradiction.
- If g is false, then $\mathfrak{A}$ does not prove it (i.e., $\mathfrak{A} \nvdash g$) since we assumed it was sound. Therefore, $\gamma(\ulcorner g \urcorner)$ is true. By completeness, $\mathfrak{A} \vdash \gamma(\ulcorner g \urcorner)$. But then, by construction, $\mathfrak{A} \vdash g$. This is a contradiction.

We conclude that $\mathfrak{A}$ cannot exist. Hence, no finitary set of laws is both sound and complete for arithmetic at the same time.

The limits of AI

Tarski's high-school problem and Gödel's incompleteness theorems are more than curiosities in the cabinet of mathematical results. They speak directly about the practice of mathematics and its limits. What significance do they hold for the working mathematician today? Recently, it has been shown that they have implications for the practice of artificial intelligence (AI).

Today the axiomatic foundations of mathematics are based on set theory rather than geometry or arithmetic. Typically, we use Zermelo–Fraenkel set theory with the axiom of choice (ZFC). For other choices, seek spiritual guidance from your nearest set theorist. Gödel's incompleteness theorems apply in this setting too.

In 1963, Paul Cohen [7] showed that the *continuum hypothesis* (CH) can neither be proved nor disproved from the axioms of ZFC, just like g can neither be proved nor disproved from $\mathfrak{A}$. The CH asks whether there exists a set whose cardinality lies strictly between that of the integers and the real numbers. It may sound like an arcane concern, too abstract to be seriously meaningful to any practical end, but it does matter. It matters in the same way that those axioms we are given in high-school limit what we can do in arithmetic.

The 2019 work by Shai Ben-David et al. [8] studies the EMX (estimating the maximum) machine learning model. The basic idea is as follows:

- You are given a class of functions (like rules or strategies).
- You get some random samples (examples).
- You want to choose a function that, on average, gives close to the best possible score over all data, not just the data you have seen.

This is a natural and important problem. It models many real-world tasks: advert selection, medical decision-making, resource allocation and more.

Ben-David et al. asked: Is there always a learning algorithm that works for a given class of functions? In other words, can we prove that some algorithm can always discover the right function (or that no algorithm can)? No. For one specific function class they constructed, its learnability actually depends on CH. If CH is true, then the class is not learnable. If CH is false, then the class is learnable. But CH is independent of ZFC, so you cannot prove it either way. This means that whether a given learning problem is solvable or not depends not on data or algorithms but on your mathematical metaphysics.

What seemed like a modest gap in high-school algebra turns out to echo a deep and general phenomenon: the impossibility of capturing all mathematical truths within any single, consistent, finitely describable system. As recent work in machine learning shows, this practical question speaks to our ability to develop the mathematical basis of AI.

References

1 Dedekind, R. (1888) *Was sind und was sollen die Zahlen?* Friedr. Vieweg & Sohn, Braunschweig.

2 Burris, S. and Lee, S. (1993) Tarski's high school identities, *Am. Math. Mon.*, vol. 100, no. 3, pp. 231–236.

3 Gurevič, R. (1990) Equational theory of positive numbers with exponentiation is not finitely axiomatizable, *Ann. Pure Appl. Log.*, vol. 49, no. 1, pp. 1–30.

4 Hilbert, D. (1950) *The Foundations of Geometry*, trans. Townsend, E.J., Open Court (original work published 1899).

5 Gentzen, G. (1969) Investigations into Logical Deduction, in *The Collected Papers of Gerhard Gentzen*, trans. Szabo, M.E., North-Holland, pp. 68–131 (original work published 1935).

6 Cantor, G. (1996) On an elementary question in the theory of manifolds, in *From Kant to Hilbert: A Source Book in the Foundations of Mathematics, Vol. 2*, ed. Ewald, W.B., Oxford University Press, pp. 920–922 (original work published 1891).

7 Cohen, P.J. (1963) The independence of the continuum hypothesis, *Proc. Natl. Acad. Sci. USA*, vol. 50, no. 6, pp. 1143–1148.

8 Ben-David, S. et al. (2019) Learnability can be undecidable, *Nat. Mach. Intell.*, vol. 1, pp. 44–48.