2002

The Continuum Problem

John Stillwell
University of San Francisco, stillwell@usfca.edu

Recommended Citation
John Stillwell. The Continuum Problem. The American Mathematical Monthly. Vol. 109, No. 3 (Mar., 2002), pp. 286-297. DOI: 10.2307/2695360

THE EVOLUTION OF...

Edited by **Abe Shenitzer and John Stillwell**

The Continuum Problem

John Stillwell

1. INTRODUCTION. In 1900, Hilbert made Cantor's continuum problem number one on his list of mathematical problems for the twentieth century. In 2000 it no longer ranked so highly, not being among the Clay Millennium Prize Problems, for example. Indeed, some mathematicians are under the impression that the continuum problem has been settled, perhaps because of the following statement by Paul Cohen in 1985:

> My personal view is that I regard the present solution of the problem as very satisfactory. I think that it is the only possible solution. It gives a feeling for what is possible and what's impossible, and in that sense I feel that one should be very satisfied. There are further problems, but they are fairly technical ones. If I were a betting man, I'd bet no one is going to come up with any other kind of solution.
>
> (Interview with Don Albers and Constance Reid, July 1985)

However, set theory has developed enormously since the time of Cohen's great results in the 1960s. Hugh Woodin [7] has recently written an update that explains why the cardinality of the continuum should still be pursued, and why the answer expected by Cantor is probably wrong. Woodin's work is highly technical, but it grows out of classical themes in the study of the continuum. The present article is a modest attempt to describe these themes, and how they have influenced set theorists today.

2. DISCRETE AND CONTINUOUS. Since ancient times, mathematicians have realized that it is difficult to reconcile the continuous with the discrete. We understand counting 1, 2, 3, ... up to arbitrarily large numbers, but do we also understand moving from 0 to 1 through the continuum of points between them? Around 450 BCE, Zeno thought not, because continuous motion involves infinity in an essential way. As he put it in his *paradox of the dichotomy*:

> There is no motion because that which is moved must arrive at the middle (of its course) before it arrives at the end.
>
> (Aristotle, *Physics*, Book VI, Ch. 9)

We know of Zeno's ideas only through Aristotle, who is trying to debunk them, but presumably the idea here is that before arriving at the end one must get halfway, and before that quarterway, and so on, hence *continuous motion involves the completion of infinitely many acts*.

The completed infinite was taboo for the ancients, and indeed for most mathematicians until the late nineteenth century. After all, isn't an infinite process one that goes on forever, and hence remains *incomplete*? If we agree, then the continuum must remain a hazy mystery, about which we can say almost nothing. Perhaps individual points can be known, and indeed by 350 BCE Eudoxus had reached essentially the

modern view that any point is uniquely determined by its position relative to the rational points. But the secret of continuity remains out of reach, as long as we reject the completed totality of points.

In 1858, Dedekind felt "overpowering dissatisfaction" with this situation, and resolved to "secure a real definition of the essence of continuity." He tells us that he succeeded on November 24, 1858, and in doing so he made the first real advance in our understanding of the continuum since Eudoxus.

For the modern mathematician, Dedekind's construction of the real number continuum $\mathbb{R}$ is profoundly simple: take the set $\mathbb{Q}$ of rationals, and *define* the irrationals to be the *gaps* in $\mathbb{Q}$ (or "cuts" as they are often called). That is, an irrational is a partition of $\mathbb{Q}$ into two sets, $\mathbb{Q}_L$ and $\mathbb{Q}_U$, such that

- each member of $\mathbb{Q}_L$ is less than all members of $\mathbb{Q}_U$,
- $\mathbb{Q}_L$ has no greatest member, $\mathbb{Q}_U$ has no least member.

Thus each individual irrational is determined by its position in the rationals, as for Eudoxus, but now we consider the *totality* $\mathbb{R}$ of rationals and irrationals, and we see that it has *no gaps*, by construction.

This definition of the continuum could not be more convincing, but it makes an irrevocable commitment to completed infinite sets: each point is determined by a set of rationals (say, the set $\mathbb{Q}_L$), and $\mathbb{R}$ itself is a set *of* such sets. Perhaps we can avoid viewing $\mathbb{Q}_L$ as a completed infinity, since it may be a set we can step through discretely like $1, 2, 3, \ldots$, but there is no way to do this for $\mathbb{R}$. This is where the modern struggle with the continuum begins, with Cantor in the 1870s.

3. COUNTABLE AND UNCOUNTABLE. The *countable* infinite sets are those that can be ordered (or "listed") in such a way that each element has only finitely many predecessors. The prototype example is the set $\mathbb{N}$ of natural numbers, whose natural ordering is such a list:

$$0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, \ldots.$$

Since any member of the set is reached in a finite number of steps (each step being addition of 1), there is no need to imagine infinitely many steps actually being completed. One is free to regard a countable set as *potentially*, but not *actually*, infinite. This is the only type of infinity considered to exist by the ancient Greeks, and also by other eminent mathematicians such as Gauss.

With a little ingenuity, many other useful infinite sets can be "counted," and hence made acceptable by these strict standards of mathematical existence. They include the integers

$$0, 1, -1, 2, -2, 3, -3, 4, -4, 5, -5, 6, -6, \ldots,$$

the positive rationals

$$\frac{1}{1}, \frac{2}{1}, \frac{1}{2}, \frac{3}{1}, \frac{2}{2}, \frac{1}{3}, \frac{4}{1}, \frac{3}{2}, \frac{2}{3}, \frac{1}{4}, \frac{5}{1}, \frac{4}{2}, \frac{3}{3}, \frac{2}{4}, \frac{1}{5}, \ldots$$

(the rule here being to list fractions for which numerator and denominator have sum 2 first, then those for which the sum is 3, then those for which the sum is 4, etc.), and all the rationals

$$0, \frac{1}{1}, -\frac{1}{1}, \frac{2}{1}, -\frac{2}{1}, \frac{1}{2}, -\frac{1}{2}, \frac{3}{1}, -\frac{3}{1}, \frac{2}{2}, -\frac{2}{2}, \frac{1}{3}, -\frac{1}{3}, \ldots.$$

A more sophisticated example, pointed out by Cantor (1874), is the set of algebraic numbers. Each algebraic number is the root of a polynomial equation $a_n x^n + a_{n-1}x^{n-1} + \cdots + a_1 x + a_0 = 0$ with integer coefficients, to which Cantor assigned the "height" $|n| + |a_n| + |a_{n-1}| + \cdots + |a_1| + |a_0|$. There are only finitely many equations with a given height and, of course, only finitely many roots of a given polynomial equation. Hence a listing of the algebraic numbers may be derived from a listing of equations according to increasing height.

These examples show that the concept of a countably infinite set is of wide scope, perhaps wide enough to raise the hope that any infinity is countable and hence merely "potentially" infinite. Cantor himself appears to have believed that he could prove $\mathbb{R}$ to be countable, and was taken aback when he discovered otherwise in late 1873. His first reaction was to draw the positive conclusion that here was a new proof that not all numbers are algebraic, and this was how he first presented uncountability to the world, in 1874.

Of course, it was not long before the uncountability of $\mathbb{R}$ was recognized as fundamental, and its importance was reflected in at least three different proofs.

- Any countable set has gaps (Cantor 1874).
  Given a list of real numbers $x_0, x_1, x_2, x_3, \ldots$, Cantor sifts through them to find a "gap": members of the list $a_0 < a_1 < a_2 < \ldots < b_2 < b_1 < b_0$ with no x_i between the a_j and b_k. Thus he generalizes the known gaps in known countable sets (such as $\sqrt{2}$ in the rationals).
- Any countable set has measure zero (Harnack 1885).
  Given a list of real numbers $x_0, x_1, x_2, x_3, \ldots$, Harnack covers x_i by an interval of length $\varepsilon/2^{i+1}$, thus covering the whole set $\{x_0, x_1, x_2, x_3, \ldots\}$ by a set of total length less than ε. Since the line $\mathbb{R}$ has infinite length, the numbers $x_0, x_1, x_2, x_3, \ldots$ make up "almost none" of $\mathbb{R}$.
- Any countable set can be diagonalized (Cantor 1891).
  Given (say) decimal expansions of real numbers $x_0, x_1, x_2, x_3, \ldots$, Cantor constructs a number x different from each x_i by making x unequal to x_i at the ith decimal place (taking care to avoid an x with two different decimal expansions).

The first proof of uncountability looks back to Dedekind's definition of the irrationals as the gaps in the rationals. It reveals a pleasant harmony between modern and ancient senses of the word "complete": if $\mathbb{R}$ is complete in the sense of having no gaps, then we must accept it as a completed infinity, because it is not countable. The second and third proofs look forward to two of the most important themes in set theory: measurability and the diagonal argument, both of which generalize to larger sets.

In particular, the diagonal argument immediately generalizes to show that any set X is smaller than its *power set* $\mathcal{P}(X)$ (the set of all subsets of X). Instead of decimal expansions, one considers *characteristic functions* of subsets of X and, if there is a subset S_x of X paired with each x in X, diagonalization gives the set $S = \{x : x \notin S_x\}$, different from each subset S_x at the element x. Thus X has more subsets S than elements x.

4. SETS OF CONTINUUM CARDINALITY. The concept of "size" for sets, implicit in the previous examples where one set was said to be "smaller" than another, is formalized by means of one-to-one mappings. Cantor said that sets A and B have the same *cardinality* or *cardinal number*, symbolized by writing $|A| = |B|$, if there is a bijection of A onto B. Thus all countably infinite sets have the cardinality of $\mathbb{N}$, by definition, and Cantor called this cardinality $\aleph_0$.

One also writes $|A| \leq |B|$ if there is an injection of A into B, and $|A| < |B|$ if $|A| \leq |B|$ but $|A| \neq |B|$. A convenient theorem (Cantor-Schroeder-Bernstein) states that, if $|A| \leq |B|$ and $|B| \leq |A|$, then $|A| = |B|$, thus avoiding explicit constructions of many bijections. To show that A and B have the same cardinality it is enough to map A one-to-one into B and B one-to-one into A. Thanks to the Cantor-Schroeder-Bernstein theorem, it is easy to show that all the following sets have the same cardinality as $\mathbb{R}$:

- Any interval—open, half-open, or closed
- The interval $[0, 1)$ (which we view as the set of binary expansions)
- The set $2^{\mathbb{N}}$ of infinite sequences of 0s and 1s

For this reason, the cardinality of $\mathbb{R}$ is denoted by $2^{\aleph_0}$. We also see that $\aleph_0 < 2^{\aleph_0}$, since there is an obvious injection of $\mathbb{N}$ into $\mathbb{R}$, but no injection in the opposite direction, by the uncountability of $\mathbb{R}$. If we interpret each infinite sequence of 0s and 1s as the characteristic function of a subset of $\mathbb{N}$, then we see that $2^{\aleph_0}$ is also the cardinality of the set $\mathcal{P}(\mathbb{N})$ of subsets of $\mathbb{N}$.

Now any infinite sequence σ of 0s and 1s can be split into a pair (σ_0, σ_1) of such sequences, consisting of its even and odd places. Conversely, σ can be reconstructed from the pair (σ_0, σ_1). This gives the result discovered by Cantor in 1877, which led him to exclaim "I see it, but I don't believe it": the plane $\mathbb{R}^2$ has the same cardinality as the line $\mathbb{R}$.

More generally, one can split σ into three, four, ..., or even infinitely many sequences (to get infinitely many, split into the subsequences of places with $1, 2, 3, \ldots$ prime factors). This shows that the following sets also have the same cardinality as $\mathbb{R}$:

- $\mathbb{R}^2, \mathbb{R}^3, \ldots, \mathbb{R}^\omega$, where $\mathbb{R}^\omega$ denotes the set of sequences $(x_0, x_1, x_3, \ldots)$ of sequences of real numbers
- The set of continuous functions $\mathbb{R} \to \mathbb{R}$ (because each such function is determined by its values $f(r)$ on the countably many rational points r, i.e., by an element of $\mathbb{R}^\omega$)

However, the general diagonal argument shows that $\mathbb{R}$ has more subsets than elements, so none of the sets just mentioned is as large as the set $\mathcal{P}(\mathbb{R})$ of all subsets of $\mathbb{R}$, or the set of all real functions, both of which have cardinality $2^{2^{\aleph_0}}$. Nevertheless, the ubiquity of the cardinality $2^{\aleph_0}$ among uncountable sets of reals led Cantor to conjecture in 1878 that it was the cardinality of *any* uncountable subset of $\mathbb{R}$. This was the first version of the continuum hypothesis.

Weak continuum hypothesis (Cantor 1878). *Any uncountable set of real numbers has cardinality* $2^{\aleph_0}$.

5. UNCOUNTABLE ORDINALS. In 1883, Cantor found another approach to uncountability, more intricate than the power set operation, but also more delicate, because it gives the *smallest* uncountable set. He generalized the idea of counting into the infinite, obtaining the sequence of *ordinal numbers*

$$0, 1, 2, 3, \ldots, \omega, \omega + 1, \omega + 2, \ldots, \omega \cdot 2, \omega \cdot 2 + 1, \omega \cdot 2 + 2, \ldots.$$

Cantor operated on the intuitive principles that

- 0 is the least ordinal,
- any ordinal α has a unique successor $\alpha + 1$,
- any set S of ordinals has a least upper bound.

In the 1920s, von Neumann formalized these principles by defining

- $0 = \{\}$ (the empty set),
- $\alpha + 1 = \alpha \cup \{\alpha\}$ (which implies that $n + 1 = \{0, 1, 2, \ldots, n\}$ for finite n),
- lub $S = \bigcup_{\beta \in S} \beta$ (which implies that $\omega = \{0, 1, 2, \ldots\}$).

Ordinal numbers are well-ordered by the membership relation $\in$, that is, they are linearly ordered and any set of them has an $\in$-least member. And they measure all well-orderings, in the sense that any well-ordered set is isomorphic to a unique ordinal.

The finite ordinals are just the natural numbers, and their least upper bound is the first infinite ordinal ω. The countable ordinals $\omega, \omega + 1, \omega + 2, \ldots, \omega \cdot 2, \ldots$ form a huge collection that Cantor called the *second number class*. This class is closed under successor and countable unions, but nevertheless it is a set, and hence it has a least upper bound, called ω_1. Clearly, ω_1 is the least uncountable ordinal, and its cardinality is called $\aleph_1$. This led Cantor to strengthen the continuum hypothesis as follows:

Strong continuum hypothesis (Cantor 1883). $2^{\aleph_0} = \aleph_1$.

6. THE AXIOM OF CHOICE. The strong continuum hypothesis is more appealing than the weak continuum hypothesis, but also less plausible, because it implies that $\mathbb{R}$ *can be well-ordered*. Given a one-to-one correspondence between the reals and the countable ordinals, the reals are well-ordered by (the order of) their corresponding ordinals.

Cantor in fact believed that any set can be well-ordered, but no such ordering is known for $\mathbb{R}$. Well-ordering of a set S implies, among other things, that there is a *choice function* for subsets of S, that is, a function f such that $f(X)$ belongs to X for each nonempty subset X of S. Namely, we can take $f(X)$ to be the least member of X, according to the well-ordering of S. Conversely, a choice function for the subsets of S can be used to well-order S.

This *well-ordering theorem* was first proved by Zermelo in 1904, and was controversial, but today it is seen as an "obvious" transfinite induction. Just as ordinary induction is based on the fact that any natural number can be reached from 0 by the successor operation, transfinite induction is based on the fact that any ordinal can be reached by the successor and least upper bound operations.

Given a choice function f for subsets of S, one transfinitely lists the members $s_0, s_1, s_2, \ldots, s_\omega, \ldots$ of S as follows:

$$\begin{aligned} s_0 &= f(S), \\ s_1 &= f(S - \{s_0\}), \\ &\vdots \\ s_\beta &= f(S - \{s_\alpha : \alpha < \beta\}), \\ &\vdots \end{aligned}$$

This process assigns ordinal subscripts to members of S "indefinitely" in the following sense: as long as the subset of S that has been assigned ordinals, $\{s_\alpha : \alpha < \beta\}$, is not all of S, we can assign an ordinal to at least one more member of S. It follows that all members of S are assigned ordinals, and S is thereby well-ordered.

Zermelo's assumption of a choice function is called the *axiom of choice* (AC), because it has no proof from other axioms of set theory. Most mathematicians accept Zermelo's axiom, because of the greater regularity it affords in many parts of mathematics. In particular, it is generally assumed that $\mathbb{R}$ can be well-ordered, and hence its cardinality is one of the alephs $\aleph_1, \aleph_2, \aleph_3, \ldots$ obtained by iterating the above construction of $\aleph_1$. However, the axiom of choice (and hence also the strong continuum hypothesis) has consequences that make $\mathbb{R}$ look *irregular* in some respects. In particular, it implies that some subsets of $\mathbb{R}$ are not *measurable*.

7. MEASURE. Since the continuum hypothesis makes a claim about all uncountable sets of reals, it is natural to explore these sets and see the extent to which they can be proved to have cardinality $2^{\aleph_0}$. Cantor made a start on this project, showing that all uncountable closed sets have cardinality $2^{\aleph_0}$. More progress was made around 1900 by the French school of analysts headed by Borel, Baire, and Lebesgue. They needed a large class of sets to satisfy the demands of integration theory.

The traditional Riemann integral $\int_a^b f(x)\,dx$ is defined for all f that are continuous on $[a, b]$, but it fails to be defined for most discontinuous functions, even for f that are monotonic limits of continuous functions f_n. The problem is that the traditional concept of area measure is not general enough to give a meaning to the "area under the graph of f," even when f is a limit of continuous functions f_n.

In 1898 Borel extended the concept of measure to all subsets of $\mathbb{R}^n$ that result from open sets by the operations of taking complements and forming countable unions. These are now known as the *Borel sets*, and each of them is assigned a unique measure by the following rules:

- The measure of a Cartesian product of open intervals is the product of the lengths of the intervals. (More generally, congruent sets have the same measure.)
- If $A \subseteq B$, and A and B have measures $\mu(A)$ and $\mu(B)$, then the measure $\mu(A - B)$ of $A - B$ is given by

$$\mu(A - B) = \mu(A) - \mu(B).$$

- If $B = A_0 \cup A_1 \cup A_2 \cup \cdots$ is a disjoint union of sets with measures $\mu(A_0)$, $\mu(A_1)$, $\mu(A_2)$, $\ldots$, respectively, then

$$\mu(B) = \mu(A_0) + \mu(A_1) + \mu(A_2) + \cdots .$$

The Borel sets obviously include all open sets (as countable unions of products of intervals), and all closed sets (as complements of open sets), but they extend much further than this. There is a natural *hierarchy* of Borel sets with ω_1 levels: the bottom level consists of the open and closed sets, and level β comprises all countable unions of sets from levels below β, for any countable ordinal β. Using a diagonal argument, Lebesgue proved in 1905 that, for each countable ordinal β, there are Borel sets at level β that do not occur at lower levels.

Thus the complexity of Borel sets strictly increases for ω_1 steps, extending measurability to sets far beyond the open and closed sets. The *Lebesgue measurable sets* go a little farther, by assigning measure zero to arbitrary subsets of Borel sets of measure zero. The corresponding *Lebesgue integral* extends integrability to a class of functions far larger than the class of continuous functions. Moreover, the Lebesgue integrable functions have better closure properties, such as closure under monotonic convergence.

The Borel sets also satisfy the weak continuum hypothesis. That is, any uncountable Borel set has cardinality $2^{\aleph_0}$. This was proved in 1915 by Aleksandrov, a member of

Luzin's group in Moscow that studied the works of the French school and extended their ideas far beyond the Borel sets.

In his 1905 paper, Lebesgue casually claimed another closure property of the Borel sets: namely, that the orthogonal projection of a Borel set is also Borel. In 1916, Luzin's student Suslin found a counterexample to this claim, and thus showed that *the projection operation leads to a larger class of sets*. The sets generated from the Borel sets by projection (and complementation, since the projection of a projection is obviously a projection) are now called the *projective sets*, and they form a hierarchy of length ω. Level 1 consists of the projections of Borel sets (called *analytic sets*) and their complements, while level $n+1$ consists of projections of sets at level n, and the complements of these projections.

In 1917, Luzin proved that analytic sets are measurable, and that they satisfy the weak continuum hypothesis. However, he was *not* able to extend this result farther into the projective hierarchy, and by 1925 he was ready to make a remarkable prophecy:

> One does not know, and one will never know, whether the projection of the complement of an analytic set (supposed uncountable) has the cardinality of the continuum, ... nor whether it is measurable.
>
> (Luzin in *Comptes rendus Acad. Sci. Paris* **180** (1925), p. 1818)

Why was Luzin so pessimistic about determining cardinality and measurability? Well, it was already known that the axiom of choice implies the existence of nonmeasurable sets. This was proved by Vitali in 1905, and Vitali's nonmeasurable set N is definable very simply from a choice function for subsets of $\mathbb{R}$.

One defines an equivalence relation $\sim$ on $\mathbb{R}$ by $x \sim y \Leftrightarrow x - y$ is rational, and lets N be a set with exactly one member from each $\sim$-equivalence class. If the elements of N are all chosen from $[0, 1]$, then it is easy to see that the circle $\mathbb{R}$ mod 1 is the disjoint union of countably many translates of N (by rational numbers). If N is measurable, then all its translates have the same Lebesgue measure $\mu(N)$ as N, and both of the assumptions $\mu(N) = 0$ and $\mu(N) > 0$ lead to contradictions.

The unknown element in this construction is, to be sure, the choice function that takes one element from each $\sim$-equivalence class. The axiom of choice says that such a function exists, but gives no information about it. However, we know that a well-ordering of $\mathbb{R}$ immediately gives a choice function, hence the complexity of the nonmeasurable set N is essentially the same as that of a well-ordering of $\mathbb{R}$, if such a thing exists. (In particular, if $2^{\aleph_0} = \aleph_1$, then $\mathbb{R}$ has a well-ordering of length ω_1, and there is a nonmeasurable set of about the same complexity as this well-ordering. This is why we said, in the previous section, that the strong continuum hypothesis makes $\mathbb{R}$ look irregular in some respects.)

Luzin evidently suspected that well-orderings of $\mathbb{R}$ could lie at a low level in the projective hierarchy, bringing with them nonmeasurable sets at the same level. He was on the right track, but set theory could not progress farther in this direction until it acquired new tools from mathematical logic.

8. GÖDEL AND COHEN. The two great Gödel theorems from 1931, the first and second *incompleteness* theorems, owe their existence to the diagonal argument, and hence are directly inspired by set theory. Indeed, the "incompleteness" to which the first theorem refers is analogous to the incompleteness of countable sets of reals.

To see why, notice that the "diagonal number" x different from the given numbers $x_0, x_1, x_2, \ldots$ can actually be *computed* from them. To be specific, we could set

$$n\text{th digit of } x = \begin{cases} 1 & \text{if the } n\text{th digit of } x_n \text{ is not 1,} \\ 2 & \text{otherwise.} \end{cases}$$

If the list $x_0, x_1, x_2, \ldots$ is computable, in the sense that the ith digit of x_j is a computable function of i and j, then x is also computable, in the sense that its nth digit is a computable function of n. It follows that *there is no computable list of all computable real numbers*. This result has implications for axiom systems, because an axiom system is supposed to produce a computable list of theorems. It means that no consistent axiom system can produce a complete list of theorems of the form "program n defines a computable real number," for the output of such an axiom system could be diagonalized to produce a (program for) a new computable real. (An inconsistent system can produce all these theorems, but only because it proves everything!)

The argument just given is not the same as Gödel's—it is somewhat informal and in fact closer to an argument discovered by Post in 1921 but not published until twenty years later—but it contains the same essential idea. Gödel could not speak about computable real numbers because in 1931 computability did not have a mathematical definition. However, it does now (since Turing in 1936), and we can assume that any axiom system for set theory is capable of expressing it. We therefore have:

Gödel's first incompleteness theorem. *For any consistent axiom system* Σ *for set theory, there is a true sentence* τ *about real numbers not proved by* Σ.

The second theorem is more subtle, but it follows from the first by examining the role of the assumption that Σ is consistent. This assumption can itself be expressed in the language of Σ, as a sentence $\mathrm{Con}(\Sigma)$. But $\mathrm{Con}(\Sigma)$ cannot be proved, as it turns out that this would yield the unprovable Gödel sentence τ. Thus we have:

Gödel's second incompleteness theorem. *If* Σ *is a consistent, sufficiently strong axiom system for set theory, then* $\mathrm{Con}(\Sigma)$ *is not provable in* Σ.

The phrase "sufficiently strong" here means strong enough to express and prove the basic properties of computation; in particular, Σ is able to express the relation of provability. Quite weak systems, even fragments of number theory, are able to do this and it is certainly true of standard systems of set theory, which are supposed to express *all* mathematical concepts.

Gödel's theorems tell us that, even if these systems can express all concepts, they cannot prove all facts—there are "gaps" in what they can prove, corresponding to the "gaps" in countable sets of reals. However, there seems to be a difference between the gaps in number theory and the gaps in set theory. No one expects the classical conjectures of number theory, such as the twin prime conjecture, to be true but unprovable. All known true but unprovable sentences of number theory originate in logic, as sentences equivalent to $\mathrm{Con}(\Sigma)$ or the like. In contrast, the unprovable sentences of set theory include those of utmost interest: the axiom of choice, the continuum hypothesis, and the existence of nonmeasurable sets.

The first step towards revealing the gaps in set theory was taken by Gödel in 1938, for the now-standard ZF (Zermelo-Fraenkel) axiom system. He showed that certain sentences could not be disproved:

Gödel's consistency theorems. *It is consistent with the ZF axioms to assume the axiom of choice, the continuum hypothesis, and the existence of nonmeasurable sets at level 2 in the projective hierarchy.*

The method by which Gödel proved these results was by modelling the ZF axioms by what he called the *constructible* sets. These are roughly the sets that have "names" when the language of ZF is enlarged by names for the ordinals. It follows easily that the universe of constructible sets satisfies the axioms of ZF, and that it is well-ordered, since the collection of "names" inherits a well-ordering (rather like alphabetical ordering) from the ordering of the ordinals. A more subtle proof shows that all constructible reals have names involving only countable ordinals, and hence that there are only $\aleph_1$ of them, so the continuum hypothesis is true in the constructible sets. Finally, the well-ordering of constructible reals turns out to be a level 2 projective set, and this gives nonmeasurable sets at the same level.

Gödel believed that the axiom of choice and the continuum hypothesis are in fact neither provable nor disprovable from the ZF axioms, but he was unable to show this much. His suspicions were finally confirmed by Cohen in 1963:

Cohen's independence theorems. *The axiom of choice and the continuum hypothesis are not provable in ZF.*

To show this, Cohen introduced a powerful new method he called "forcing," which takes a small model of ZF and adds elements in such a way that the axioms remain satisfied, but other specific sentences are violated. In particular, he showed that it is possible to simultaneously satisfy the Zermelo-Fraenkel axioms and the axiom of choice, *and* admit almost any value for $2^{\aleph_0}$. For example, the values $\aleph_2$, $\aleph_3$, and $\aleph_{\omega+1}$ are all possible, so the continuum hypothesis can be violated in many different ways.

Taken together, the results of Gödel and Cohen show that set theory is highly incomplete, because it does not answer some of the most natural questions about sets. As early as 1947, Gödel anticipated this situation:

> One may on good reason suspect that the role of the continuum problem in set theory will be this, that it will eventually lead to the discovery of new axioms which will make it possible to disprove Cantor's conjecture.
>
> (*American Mathematical Monthly*, **54** (1947), p. 524)

And Cohen initially had a similar view [**2**, p. 151]:

> A point of view which the author feels may eventually come to be accepted is that CH [the continuum hypothesis] is *obviously* false.... $\aleph_1$ is the set of countable ordinals and this is merely a special and the simplest way of generating a higher cardinal.... The set C [the continuum] is, in contrast, generated by a totally new and powerful principle, namely the Power Set Axiom.

As we saw at the beginning of this article, Cohen felt more satisfied with the situation after another twenty years, and was content to leave the continuum in limbo. However, by this time the search for new axioms was well under way.

9. LARGE CARDINAL AXIOMS. The ZF axioms for set theory state roughly that

- $\mathbb{N}$ is a set,
- sets result from certain other operations, the most important of which are *power* (taking all subsets of a set) and *replacement* (taking the range of a function whose domain is a set).

Because of this, ZF can be modelled by any set that contains $\mathbb{N}$ and is closed under power and replacement. Such sets are called *strongly inaccessible*, and it is not hard

to believe they exist, if one believes in the axioms of ZF. However, strongly inaccessible sets cannot be *proved* (by ZF) to exist, since by Gödel's second incompleteness theorem their existence implies Con(ZF).

Thus the existence of strongly inaccessible sets is a new axiom, of a type called an *axiom of infinity*, or *large cardinal axiom*, because it claims there is a set larger than any that can be proved to exist by the other axioms of ZF. Many other large cardinal axioms have now been studied, and the size of the corresponding cardinals nicely measures the strength of the conclusions one can draw from their existence. Typically, one proves that

$$\text{Con(ZF + axiom A)} \quad \Rightarrow \quad \text{Con(ZF + axiom B)}$$

—the consistency of ZF + B *relative* to the consistency of ZF + A. If the converse is also proved, then one has *equiconsistency* of ZF + A and ZF + B.

For example, there is an interesting interplay between measurability assumptions and certain large cardinal axioms.

We know from Vitali's example that not all sets of reals can be Lebesgue measurable if the axiom of choice holds; however, if we drop translation invariance (keeping only countable additivity), the conflict with the axiom of choice disappears. We can assume that all sets of reals are measurable in this weaker sense, but the continuum then becomes very large—much larger than the smallest strongly inaccessible set. This was discovered by Ulam in 1930, and the cardinals whose subsets are all measurable are now called *measurable cardinals*. Thus

$$\begin{aligned} &\text{Con(ZF + axiom of choice + "a measurable cardinal exists")} \\ \Rightarrow\ &\text{Con(ZF + axiom of choice + "a strong inaccessible exists").} \end{aligned}$$

If we drop the axiom of choice, on the other hand, Solovay showed in 1970 that it is consistent to assume that all sets of reals are Lebesgue measurable, provided we also assume Con(ZF + axiom of choice + "a strong inaccessible cardinal exists"). In fact, ZF + "all sets of reals are Lebesgue measurable" is equiconsistent with ZF + axiom of choice + "a strong inaccessible exists," because in 1984 Shelah proved the opposite direction of Solovay's relative consistency result.

Another important consequence of measurable cardinals was proved by Scott in 1960: if there is a measurable cardinal, then not every set is constructible. This puts measurable cardinals in conflict with the concept of constructibility used by Gödel to prove the consistency of the continuum hypothesis; however, measurable cardinals do *not* contradict the continuum hypothesis itself. Many models are now known to satisfy the continuum hypothesis, and it appears that no large cardinal axiom alone will contradict it.

Nevertheless, there is another way in which large cardinal axioms can illuminate the continuum hypothesis, and this is the subject of our last section.

10. DETERMINACY. A new direction in the study of sets of reals was initiated by Polish mathematicians in the 1920s—though for a long time it remained little known and was not considered important—the theory of infinite games. In 1925 Steinhaus associated a 2-person game G_A with each set A contained in the unit interval as follows. Players I and II alternately choose binary digits of a real number x, for ω steps. At the end of play, I wins if x lies in A; otherwise II wins. If one of the players has a winning strategy for this game, then the set A is called *determined*.

For example, the set A of irrational numbers in [0, 1] is a determined set, because player I can always win by playing a nonperiodic sequence of digits, for instance,

This is a winning strategy because the real number x produced has a nonperiodic binary expansion, and hence is irrational, no matter what digits II inserts between the successive digits of I's sequence.

However, not all sets A of reals are determined, at least if we assume the axiom of choice. A counterexample was found by Banach and Mazur, also in 1925. In fact, until the 1970s, very few sets were known to be determined, and the concept of determinacy was kept alive only by connections with other concepts of set theory. To formalize these connections, Steinhaus and Mycielski in 1962 proposed the *axiom of determinacy* (AD): *every set of reals is determined.*

Set theorists find AD hard to believe, since it contradicts the axiom of choice. But it does not appear to contradict the ZF axioms, and its consequences are spectacular. In 1964 Mycielski proved that

- AD $\Rightarrow$ every set of reals is Lebesgue measurable;
- AD $\Rightarrow$ the weak continuum hypothesis.

Moreover, if one rejects determinacy of all sets of reals, one can restore the axiom of choice and still derive Lebesgue measurability and the weak continuum hypothesis for large classes of sets. For example, if one assumes *projective determinacy*—that every projective set is determined—then every projective set is Lebesgue measurable, and every uncountable projective set has cardinality $2^{\aleph_0}$. To many set theorists, this gives the best of both worlds; a well-behaved projective world, and the axiom of choice everywhere. A burning question of the 1970s and 1980s was therefore: How reasonable is projective determinacy?

Before the 1970s, determinacy was known only for the first few levels of the Borel hierarchy. In 1975, D. A. Martin made a remarkable breakthrough by proving determinacy for all Borel sets. His proof used the full resources of ZF, but not any large cardinal assumptions. Proving determinacy for larger classes of sets (or, at least, the consistency of *assuming* determinacy for these classes) does depend on large cardinal axioms, and the larger the class, the larger the cardinal required.

Already in 1970, Martin had proved determinacy for analytic sets assuming the existence of a measurable cardinal, and in 1987 he, Steel, and Woodin established the large cardinal strength of projective determinacy: it is equivalent to the existence of certain large cardinals now called *Woodin cardinals*. These cardinals are larger than measurable cardinals, though not as large as some others that have been considered. Their definition may be found in Kanamori [**4**], along with the story of large cardinals in general.

None of these results settle the continuum hypothesis, but they change the face of set theory by suggesting new axioms, such as projective determinacy. Woodin has now written a 900-page book (and he has more books on the way!) explaining how these new axioms may be expected to fill the glaring gaps in the theory of real numbers, so that only Gödel sentences and consistency statements remain unprovable. In particular, the new axioms imply that $2^{\aleph_0} = \aleph_2$, hence that *the continuum hypothesis is false*.

Modern set theory is a highly intricate subject, and no doubt it will be a long time before working mathematicians are prepared to accept new axioms, especially when the axioms cannot even be properly described in an article of this size. However, it seems to me that all mathematicians should be curious about these developments, and I urge readers to take the next step, which is to read Woodin's own introduction to his program in [**7**].

1. D. J. Albers, G. L. Alexanderson, and C. Reid, eds., *More Mathematical People*, Academic Press, San Diego, 1990.
2. P. Cohen, *Set Theory and the Continuum Hypothesis*, W. A. Benjamin, Inc., New York, 1966.
3. S. Feferman et al., eds., *Kurt Gödel Collected Works*, vol. 2, Oxford University Press, New York, 1990.
4. A. Kanamori, *The Higher Infinite*, Springer-Verlag, New York, 1994.
5. A. Kanamori, The mathematical development of set theory from Cantor to Cohen, *Bull. Symb. Logic* **2** (1996) 1–71.
6. G. H. Moore, *Zermelo's Axiom of Choice. Its Origins, Development, and Influence*, Springer-Verlag, New York, 1982.
7. W. H. Woodin, The Continuum Hypothesis, Parts I and II, *Notices Amer. Math. Soc.* **48** (2001) 567–576, 681–690.