

# The diagonal argument—a study of cases

ZVONIMIR ŠIKIĆ

Zagreb, Croatia

To W. V. O. Quine from whom I learned set theory and its logic.

## 1

The diagonal argument entered mathematics with Cantor. The argument is one of the corner-stones of his theory of infinity. Cantor asserted that the actual infinite is mathematically manageable like the finite and is to be treated this way *as far as possible*. Hence, he moved the unmanageable and mathematically undetermined absolute infinite far away from the finite. The whole realm of mathematically manageable infinities was inserted between the finite and the absolute infinite. This is the realm of the transfinite. But what is the structural difference between the transfinite and the absolute? *The transfinite is increasable while the absolute is not*. The diagonal argument was used by Cantor to prove that there are increasable infinities, or transfinities.

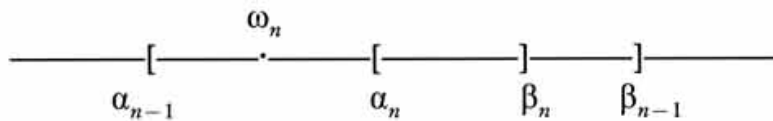
Cantor did not use the diagonal argument in his first proof (Cantor, 1874) of the increasability of  $\aleph_0$ , in which the non-denumerability of the real numbers was proved. Assuming that the real numbers were denumerable, it followed that they could be arranged in the sequence:

$$\omega_1, \omega_2, \omega_3, \dots, \omega_n, \dots; \quad n \in N. \quad (1)$$

Cantor went on to prove that it was possible, given any interval  $[\alpha, \beta]$ , to find a real number inside the interval that failed to be an element of (1). He picked the first two numbers from (1) lying within the interval  $[\alpha, \beta]$ . Denoted  $\alpha', \beta'$ , respectively, these were used to constitute another interval  $[\alpha', \beta']$ . It followed that  $[\alpha', \beta']$  missed all the elements which preceded it in (1). Proceeding analogously, Cantor produced a sequence of nested intervals  $[\alpha, \beta] \supseteq [\alpha', \beta'] \supseteq [\alpha'', \beta''] \supseteq \dots$ , such that each of the intervals missed all the elements which preceded it in (1). It followed that  $[\alpha^\infty, \beta^\infty]$ , which is the intersection of all the intervals, and which is non-empty because of the continuity of the real line, missed all the elements of (1). The contradiction established the proof: the real numbers were not denumerable. The infinity of natural numbers is not absolute, because it is increasable.<sup>1</sup>

A slight modification of Cantor's proof leads up to the diagonal argument. Let  $[\alpha_1, \beta_1]$  be a subinterval of  $[\alpha, \beta]$  which does not contain  $\omega_1$ . Let  $[\alpha_2, \beta_2]$  be a subinterval of  $[\alpha_1, \beta_1]$

which does not contain  $\omega_2$  (then it does not contain  $\omega_1$ , either). Proceedings analogously we produce  $[\alpha_n, \beta_n]$  which contains none of the first  $n$  elements of (1).



The non-empty intersection  $[\alpha_\infty, \beta_\infty] = \bigcap_{1 \leq n < \infty} [\alpha_n, \beta_n]$  contains none of the elements of (1) and that is what we wanted to prove. If we start with  $[\alpha, \beta] = [0, 1]$  and pick  $[\alpha_1, \beta_1]$  (which does not contain  $\omega_1$ ) as one of the decimal intervals  $[0, 0.1], [0.1, 0.2], \dots, [0.9, 1]$ , then the decimal expansions of  $\omega_1$  and  $\alpha_1$  differ at their first decimal places. Analogously, the decimal expansions of  $\omega_2$  and corresponding  $\alpha_2$  differ at their second decimal places,  $\omega_3$  and  $\alpha_3$  at their third decimal places, and so on:

$$\begin{array}{lll} \omega_1 = 0. d_{11} d_{12} d_{13} \dots & \alpha_1 = 0. d_1 & , \quad d_1 \neq d_{11} \\ \omega_2 = 0. d_{21} d_{22} d_{23} \dots & \alpha_2 = 0. d_1 d_2 & , \quad d_2 \neq d_{22} \\ \omega_3 = 0. d_{31} d_{32} d_{33} \dots & \alpha_3 = 0. d_1 d_2 d_3 & , \quad d_3 \neq d_{33} \\ \vdots & \vdots & \vdots \end{array}$$

The decimal expansions of  $\alpha_\infty$  and  $\alpha_n$  agree at their first  $n$  decimal places. It follows that the decimal expansion of  $\alpha_\infty$  is constructed so as to differ at each decimal place from the *diagonal expansion*  $d$ . The decimal expansion of  $\alpha_\infty$  is antidiagonal:

$$\begin{array}{l} \alpha_\infty = 0. d_1 d_2 d_3 d_4 \dots \\ \quad \quad \quad \uparrow \uparrow \uparrow \uparrow \\ d = 0. d_{11} d_{22} d_{33} d_{44} \dots \end{array}$$

Hence, there are more decimal expansions than natural numbers or, what is the same, there are more functions from  $N$  to  $\{0, 1, \dots, 9\}$  than natural numbers. We do not need real numbers to prove the increasability of the infinity of natural numbers.

Although there is no historical evidence that this particular chain of ideas led Cantor from his 1874–geometric proof to his 1891–diagonal proof, it is quite a natural chain. Whatever the case may be, it is the new diagonal proof which is the most important. It is independent of the geometric considerations of the continuity of real numbers. In Cantor’s own words: ‘*Es lässt sich aber von jenem Satze ein viel einfacherer Beweis liefern, der unabhängig von der Betrachtung der Irrationalzahlen ist*’ (Cantor, 1891). The infinity of natural numbers is increasable because there are many more functions from  $N$  to  $N$  than elements in  $N$ . Cantor used the principle of diagonalization to prove this and what is more important, the principle can be extended directly to the general theorem, that the powers of well-defined classes have no maximum or, what is the same, that in place of any given class  $L$  another class  $M$  can be placed which is of greater power than  $L$ . ‘Dieser Beweis erscheint nicht nur wegen seiner grossen Einfachkeit, sondern namentlich auch aus dem Grunde bemerkenswert, weil das darin befolgte Prinzip sich ohne weiteres auf den allgemeinen Satz ausdehnen lässt, dass die Mächtigkeiten wohldefinierter Mannigfaltigkeiten kein Maximum haben oder, was dasselbe ist, dass jeder gegebenen Mannigfaltigkeiten  $L$  eine andere  $M$  an die Seite gestellt werden kann, welche von stärkerer Mächtigkeit ist als  $L$ ’ (Cantor, 1981). The paradise of increasable infinities, or transfinities has been created.<sup>2</sup>

Let us analyse Cantor's diagonal proof in detail. Assuming that there are only denumerably many functions from  $N$  to  $N$ , or that it is possible to index all of them with natural numbers, it follows that all the functions could be arranged in the sequence

$$f_1, f_2, f_3, \dots, f_n, \dots; \quad n \in N. \tag{2}$$

This sequence of functions can be visualized as an infinite rectangular array of their values.

	1	2	3	
$f_1$	$f_1(1)$	$f_1(2)$	$f_1(3)$	$\dots$
$f_2$	$f_2(1)$	$f_2(2)$	$f_2(3)$	$\dots$
$f_3$	$f_3(1)$	$f_3(2)$	$f_3(3)$	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	

The entries in the diagonal of the array (upper left to lower right) form the sequence of values of the diagonal function,  $\delta(n) = f_n(n)$ . The antidiagonal function  $\alpha$  is defined in such a way that it differs from this diagonal function at every place:

$$\alpha(n) = \delta(n) + 1 = f_n(n) + 1, \text{ for every } n \in N. \tag{3}$$

Then the antidiagonal sequence differs from every row of our array, and so the antidiagonal function differs from every function in our sequence (2). This is a contradiction, because it was assumed that every function belongs to (2). More formally, if  $\alpha$  is one of the functions from (2) then  $\alpha = f_k$ , for some  $k \in N$ . It follows that

$$\alpha(k) = f_k(k),$$

but on the other hand it follows from (3) that

$$\alpha(k) = f_k(k) + 1.$$

From this contradiction Cantor concluded that there are more functions from  $N$  to  $N$  than natural numbers. The infinity of natural numbers is increasable.

Analogously, Cantor started with any collection  $K$  and proved that there are more functions from  $K$  to  $K$  than members of  $K$ . Namely, if it is possible to index all the functions with members of  $K$  then each of the functions is of the form  $f_x$ , for some  $x \in K$ . The antidiagonal  $\alpha$  is defined as always:  $\alpha(x) = f_x(x) + 1$ , for every  $x \in K$ . From the assumption that there is  $x \in K$  such that  $f_x = \alpha$ , a contradiction follows:  $f_x(x) = \alpha(x) = f_x(x) + 1$ . Cantor concluded that there are more functions from  $K$  to  $K$  than members of  $K$ .

Of course, if a contradiction is derived (at least) one of the assumptions is to be refuted. The question is, which one is to be blamed for the contradiction. Cantor's diagonal argument has at least three assumptions:

1. The sequence (2) contains all functions from  $N$  to  $N$ .
2. The antidiagonal  $\alpha$  is a well-defined function from  $N$  to  $N$  and, of course.
3. The elementary classical logic is valid.

Cantor did not question (2) and (3) and so he refuted (1).

## 2

This is not the only possibility. If we think about functions from  $N$  to  $N$  as finitely describable rules, as did (Kleene, 1936) and (Turing, 1936), then we are not ready to question (1). Namely, there are only denumerably many finite strings of symbols which we use to describe the rule-functions. It is not important here to specify the descriptions. They could be Turing machines or abacus programs or something else. We just call them programs, and think about a function as computable if it is computable by one of our programs. If we try to rederive the contradiction in this setting, we have to start with a complete enumeration of programs for all computable functions. If

$$p_1, p_2, p_3, \dots, p_n, \dots; \quad n \in N \quad (4)$$

is such a complete enumeration, then one possible program for the antidiagonal function  $\alpha$  is

$$\alpha(n) = p_n(n) + 1.$$

'To compute  $\alpha(n)$  take  $p_n$ , apply it to  $n$  and add 1'. Our list of programs is complete, hence  $\alpha$  could be computed with some  $p_k$ . It seems that we have the contradiction again:

$$p_k(k) = \alpha(k) = p_k(k) + 1. \quad (5)$$

But there is no contradiction here. If a program  $p$ , for some computable function  $f$ , is applied to a natural number  $n$ , the computation will certainly start, but it is possible that it will never halt. In such a case  $f(n)$  is undefined,  $f(n) = \infty$ . It is possible that  $\alpha(k) = \infty$  and there is no contradiction in (5): if  $p_k$  applied to  $k$  never halts then of course,  $p_k + 1$  applied to  $k$  never halts.

The computable functions do not need to be defined everywhere. They are partial functions. Perhaps, we could rederive the contradiction with the slightly different antidiagonal function  $\alpha_1$  which is an everywhere defined total function:

$$\alpha_1(n) = \begin{cases} 1 & \text{if } p_n(n) = \infty \\ p_n(n) + 1 & \text{otherwise.} \end{cases} \quad (6)$$

If  $\alpha_1$  could be computed with some  $p_k$  then

$$1 = \alpha_1(k) = \infty$$

if  $p_k$  applied to  $k$  never halts, or

$$p_k(k) = \alpha_1(k) = p_k(k) + 1$$

if  $p_k$  applied to  $k$  halts. Contradiction in both cases. Hence,  $\alpha_1$  is not computable. But what is wrong with (6) as a program for  $\alpha_1$ : 'To compute  $\alpha_1(n)$  take  $p_n$ , apply it to  $n$ , and if it halts add 1. If it does not halt take 1 as  $\alpha_1(n)$ '. If we could decide whether a program will stop when applied to an argument, (6) would be a program for  $\alpha_1$ . Since there is no such program, it follows that there is no program which could decide whether a program will stop when applied to an argument. The halting problem is undecidable.

Another possibility is to throw out from (4) all the programs which do not compute total functions.

$$\begin{array}{cccccccc} p_1, & p_2, & p_3, & p_4, & p_5, & p_6, & p_7, & \dots ; \\ \downarrow & \downarrow & & \downarrow & & \downarrow & & \\ out & out & & out & & out & & \end{array} \quad (7)$$

Removing these programs, the complete list of programs for total functions remains:

$$q_1(=p_3), q_2(=p_5), q_3(=p_7), \dots$$

This sequence generates its own antidiagonal function

$$\alpha_2(n) = q_n(n) + 1, \tag{8}$$

which is total. If it were computable, some  $q_k$  would compute it, what is contradictory:

$$q_k(k) = \alpha_2(k) = q_k(k) + 1.$$

Hence,  $\alpha_2$  is not computable. But what is wrong with (8) as a program for  $\alpha_2$ : 'To compute  $\alpha_2(n)$  take  $q_n$ , apply it to  $n$  and add 1'? This would be a program for  $\alpha_2$  if we could construct the sequence of  $q$ -s, by throwing out programs of non-total functions, cf. (7). Since there is no such program, it follows that there is no program which could decide whether  $p_n$  is a program for a total function or not. The class of programs for total functions is undecidable. If programs for total functions are simply called rules, it follows that there is no rule to decide what is a rule. (The same result could be derived using the antidiagonal function

$$\alpha_3(n) = \begin{cases} p_n(n) + 1 & \text{if } p_n \text{ computed total function} \\ 1 & \text{otherwise.} \end{cases}$$

We just have to argue as we did with  $\alpha_1$ )

Rice proved (Rice, 1953) that the class of all programs which compute functions from a given collection of computable functions is computable if and only if the collection of functions is trivial, i.e. either empty or containing all computable functions. Rice's theorem is quite powerful, since it incorporates a number of undecidability results. Its content is that any non-trivial functional property of programs is undecidable. Thus, undecidability proofs of given properties are reduced to proofs of non-triviality, which are usually immediate. For example, the following problems are undecidable: being total (as we have seen), being onto, being completely undefined, being defined for a given fixed argument, having a given number as value, and so on. Using a more sophisticated application of the diagonal argument we prove Rice's celebrated theorem.

If (4) is a complete enumeration of programs for computable functions of one argument, consider the program  $p$  which computes the function  $\gamma(m, n)$  of two arguments: 'Take  $p_m$  and apply it to  $m$ . If  $p_m$  halts at  $m$  take  $p_{p_m(m)}$  and apply it to  $n$ '. Hence,

$$\gamma(m, n) = p_{p_m(m)}(n), \quad \text{if } p_m(m) < \infty,$$

and  $\gamma(m, n) = \infty$  otherwise. If we execute our program  $p$  with fixed  $m$  we get a program for  $\gamma(m, n)$  with fixed  $m$ . This is a program for a computable function of one argument  $n$ , so it is identical to some program  $p_{d(m)}$  in our list. Hence, for each  $m$  we may compute  $d(m)$  such that  $p_{d(m)}$  is the program which computes a corresponding function of one argument, i.e.

$$\gamma(m, n) = p_{d(m)}(n).$$

Note that  $d(m)$  is a total computable function. If  $f$  is any other total computable function then  $f(d(m))$  is also a total computable function. Hence, it is computable with some program  $p_k$  from our list, i.e.

$$f(d(m)) = p_k(m).$$

It follows that

$$p_{d(k)}(n) = \gamma(k, n) = p_{p_k(k)}(n) = p_{f(d(k))}(n),$$

because  $p_k(k) < \infty$  (note that  $p_k$  is total). This is the content of the famous Kleene's fixed point theorem (Kleene, 1938): For any total computable  $f$  there is an index  $j$  such that  $p_j$  computes the same function as  $p_{f(j)}$ . (We proved that  $d(k)$  is such a  $j$ .)

Now, let us suppose that  $\kappa$  is a non-trivial collection of computable functions. Then, there exist programs  $p_a$  and  $p_b$  such that  $p_a$  computes a function in  $\kappa$ , and  $p_b$  computes a function which is not in  $\kappa$ , i.e.

$$p_a \in \kappa \text{ and } p_b \notin \kappa.$$

We define function  $g$  such that

$$g(n) = \begin{cases} a & \text{if } p_n \notin \kappa \\ b & \text{if } p_n \in \kappa. \end{cases}$$

It is easy to prove that  $g$  is not computable. Namely,  $p_n \in \kappa \rightarrow g(n) = b \rightarrow p_{g(n)} \notin \kappa$  and  $p_n \notin \kappa \rightarrow g(n) = a \rightarrow p_{g(n)} \in \kappa$ , which contradicts Kleene's fixed point theorem. But  $g$  would be computable if  $p_n \in \kappa$  were decidable. Hence,  $p_n \in \kappa$  is not decidable.

### 3

The considerations in 2 demonstrate quite clearly in what sense the diagonal argument is the starting point of computability theories. We return now to set theory.<sup>3</sup> Cantor's proof that there are more functions from  $K$  to  $K$  than members of  $K$  is easy to transform to the proof that there are more functions from  $K$  to  $\{0, 1\}$  than members of  $K$ . The antidiagonal  $\alpha$  is defined as follows:

$$\alpha(x) = \begin{cases} 1 & \text{if } f_x(x) = 0 \\ 0 & \text{if } f_x(x) = 1. \end{cases}$$

Since, there is a natural correspondence between functions  $f$  from  $K$  to  $\{0, 1\}$  and subclasses  $F$  of  $K$ :

$$f(x) = \begin{cases} 1 & \text{if } x \in F \\ 0 & \text{if } x \notin F, \end{cases}$$

it follows that there are more subclasses of  $K$  than members of  $K$ , i.e.  $P(K) > K$ . It is instructive to repeat the proof directly with subclasses. Let us suppose that it is possible to index all the subclasses of  $K$  with members of  $K$ , i.e. that there is an (index) map  $j$  that maps members of  $K$  onto its subclasses (so that each subclass is a value of the map). This correspondence  $j$  can be visualized as an rectangular array:

	$j$	$a$	$b$	$c$	$\dots$
$j(a) =$	$A$	1	0	1	$\dots$
$j(b) =$	$B$	0	0	1	$\dots$
$j(c) =$	$C$	0	1	1	$\dots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

Each row determines (the membership in) the respective subclass; in this particular case  $a \in A, b \notin A, c \in A, \dots$ . The diagonal of the array determines the diagonal subclass  $D$ . In this particular case  $a \in D, b \notin D, c \in D, \dots$ . Generally,  $D = \{x \in K: x \in j(x)\}$ . The antidiagonal subset  $\bar{D}$  is the complement of  $D$  in  $K$ . Hence

$$\bar{D} = \{x \in K: x \notin j(x)\}. \quad (9)$$

It follows, for any  $x \in K$ , that

$$x \in \bar{D} \leftrightarrow x \notin j(x).$$

If  $\bar{D}$  corresponds to  $\bar{d}$ , i.e.  $j(\bar{d}) = \bar{D}$ , then the following contradiction follows:

$$\bar{d} \in \bar{D} \leftrightarrow \bar{d} \notin j(\bar{d}) \leftrightarrow \bar{d} \notin \bar{D}.$$

Cantor concluded that there is no  $j$  that maps members of  $K$  onto its subclasses, i.e. there are more subclasses than elements.

Russell believed that there is an error in Cantor's proof. On 8 December 1900 he corresponded with Louis Couturat about the supposed error (Moore & Garciadiego, 1981):

I have discovered a mistake in Cantor, who maintains that there is no largest cardinal number. But the number of [the universe of all] classes is the largest number. The best of Cantor's proofs to the contrary can be found in *Jahresb. d. deutschen. Math. Ver'g.*, I, 1892, pp. 75–78 [Cantor, 1891]. In effect it amounts to showing that, if  $U$  is a class whose number is  $\alpha$ , the number of classes included in  $U$  (which is  $2^\alpha$ ) is larger than  $\alpha$ . The proof presupposes that there are classes included in  $U$  which are not members of  $U$ ; but if  $U =$  class [of all classes], that is false: [for] every class of classes is a class.

It was clear to Russell that in the universe of all classes there are not more subclasses than elements, because each subclass of the universe is also an element of the universe. Hence, there is a trivial correspondence between subclasses of the universe and its members: each subclass corresponds to itself. Russell did not try, at that time, to specify the mistake. He gave a counter-example to Cantor's proof and concluded that there had to be a mistake in it.

Replying to Russell's letter on 3 January 1902, Couturat questioned Russell's class of all classes (Moore & Garciadiego, 1981): 'I wonder whether one can consider the class of all possible classes without some sort of contradiction'. Russell and Couturat did not know that Zermelo had already used the diagonal argument to prove the inconsistency of the universe of all classes. He derived the so-called Russell's paradox from the supposed existence of the universe of all classes. He did not publish the proof, but later he claimed to have found the paradox independently of Russell (Zermelo, 1908): 'I had however discovered this antinomy myself, independently of Russell, and informed Prof. Hilbert, among others, of it even before 1903' (when it was published by Russell). Indeed, Hilbert wrote to Frege on 7 November 1903 that Zermelo found the paradox 3–4 years ago (Rang & Thomas, 1981). Fortunately, we know that Husserl was one of those colleagues 'among others', because a note is found in his *Nachlass*, which contains a detailed exposition of Zermelo's proof (Rang & Thomas, 1981):

Zermelo informs me (16 April 1902) concerning p. 272 of my Schröder reviews. In the issue, not in the method of proof Schröder was right, namely: A class  $M$ , which contains each of its subsets  $m, m', \dots$  as elements, is an inconsistent class, i.e. such a class, if at all treated as a class, leads to contradictions. PROOF. We consider those subclasses  $m$  which do not contain themselves as elements. ( $M$  contains as elements

each of its subclasses; hence subclasses of  $M$  will also contain certain subclasses as elements, themselves not [being] elements, and now we consider just those subclasses  $m$ , which may perhaps contain other subclasses, but not themselves as elements.) These constitute in their totality a class  $m_o$  (i.e. the class of all subclasses of  $M$  which do not contain themselves as elements), and now I prove of  $M_o$  [the following contradiction,

- (1) that it does not contain itself as an element,
- (2) that it contains itself as an element.

Concerning (1):  $M_o$ , being a subclass of  $M$ , is itself an element of  $M$ , but not an element of  $M_o$ . For, otherwise,  $M_o$  would contain as an element a subclass of  $M$  (namely,  $M_o$  itself) which contains itself as an element, and that would contradict the notion of  $M_o$ .

Concerning (2): Hence  $M_o$  itself is a subclass of  $M$  which does not contain itself as an element. Thus it must be an element of  $M_o$ .

Of course, a class with a definition as  $M$  is the class of all classes.<sup>4</sup>

In place of Cantor's  $K$  Zermelo took a class  $M$  which contains each of its subclasses as elements. Hence, he could take identity correspondence  $i$  instead of Cantor's  $j$ . Zermelo's antidiagonal  $\bar{D} = \{x \in M: x \notin i(x)\} = \{x \in M: x \notin x\}$  implies a contradiction, because  $\bar{D} \in \bar{D} \leftrightarrow \bar{D} \notin \bar{D}$ , or equivalently  $\bar{D} \in \bar{D} \& \bar{D} \notin \bar{D}$ . Zermelo concluded that a class  $M$  which contains each of its subclasses as elements is an inconsistent (mathematically unmanageable) multiplicity. In particular, the universe of all classes, which should contain each of its subclasses as elements, is an inconsistent, mathematically unmanageable multiplicity. From Cantor's point of view it is a trivial fact because it is evident that the unincreasable universe of all classes is the Absolute infinite, which is mathematically unmanageable. As Hilbert wrote to Frege, in the above-mentioned letter, the contradiction was well known to those in Göttingen, 'as were the other even more convincing contradictions' (Rang & Thomas, 1981).

This was not at all evident to the authors like Husserl or Russell who considered the universe of all classes as a completely legitimate class. Husserl's criticism of Schröder's alleged proof of the inconsistency of the universal class led up to the Zermelo–Husserl correspondence in which Zermelo warned Husserl that 'Schröder is right, not in the method of proof, but in the issue'. Russell, as we have seen, used the universe of all classes as a counter-example to Cantor's proof. When he finally tried to specify the mistake in Cantor's proof he applied Cantor's diagonal argument to the universe of all classes  $U$  with the identity correspondence  $i$  in place of Cantor's  $K$  and  $j$ . Russell's antidiagonal  $R = \{x \in U: x \notin x\}$  could be written down as

$$R = \{x: x \notin x\},$$

because every  $x$  belongs to the universe, and the following contradiction follows:

$$R \in R \leftrightarrow R \notin R.$$

For Russell it was not only contradiction which refutes one of its assumptions. It was a genuine paradox. [We distinguish here between the terms paradox and contradiction. A contradiction is just an inconsistent proposition (which warns us that some of its assumptions are false). On the other hand, a paradox is an argument which ends in a contradiction although all of its assumptions are acceptable. A paradox forces us to give up an assumption which we previously accepted as correct.] For Cantor, Zermelo, Hilbert and other 'Cantorians' there is nothing paradoxical in the contradictions of the universe of all classes, or the class of all ordinals, etc. These were just new, and usually uninteresting, confirmations that the Absolute infinite is inconsistent and mathematically unmanageable.<sup>5</sup> But

Russell accepted the class of all classes as a legitimate class and, even more, he accepted the principle of abstraction as a legitimate mode of generating classes. This provides that, given any condition  $C$  upon  $x$ , there is a class  $\{x: C(x)\}$  whose members are just those classes  $x$  such that  $C(x)$ , i.e. each condition or property has an extension. (Note that the principle of abstraction implies the existence of the universe of all classes:  $U = \{x: x = x\}$ .) The principle of abstraction was crucial for the logicist reduction of mathematics to logic. It was the corner-stone of Frege's *Grundgesetze* and Russell's *Principles of Mathematics*. Here we have the real paradox which threatened the whole system of Frege and Russell. Frege surrendered. Russell gave up the principle of abstraction and created his theory of types. The classes are stratified into so-called types, and the rule is imposed, that  $x \in y$  is a meaningful proposition only if the values of  $y$  are of next higher type than those of  $x$ ; otherwise  $x \in y$  is reckoned as meaningless. An expression will hence be reckoned as meaningful by the theory of types only if there is a way of so assigning types to the variables as to conform to this requirement. Such an expression is called stratified. As a consequence, the principle of abstraction is restricted since the unstratified conditions  $C(x)$  are not allowed to enter it. In particular  $x \notin x$  is not allowed. Continuing this stream of ideas, Quine gave up the (general) principle of abstraction in a simpler and more relaxed way in his 'New foundations for mathematical logic' (Quine, 1937):

Whereas the theory of types avoids the contradictions by excluding unstratified formulas from the language altogether, we might gain the same end by continuing to countenance unstratified formulas but simply limiting R3 [the principle of abstraction] explicitly to stratified formulas. Under this method we abandon the hierarchy of types, and think of the variables as unrestricted in range.

In 'Cantor's theorem and paradoxical classes' (Šijić, 1986) Cantor's diagonal argument is used to suggest an even more relaxed limitation of the principle of abstraction.

We summarize the whole discussion of the universal class and the principle of abstraction. Using the diagonal argument a contradiction is derived from the following assumptions:

1. The universe  $U$  of all classes is a consistent class, and the identity correspondence  $i$  is a natural correspondence between its subclasses and its elements, because each subclass of  $U$  is at the same time its element.
2. Given any condition  $C$  upon members  $x$  of class  $K$ , there is a class  $\{x \in K: c(x)\}$  whose member are just those members  $x$  of  $K$  such that  $C(x)$ .

(Taken together the assumptions are equivalent to the (general) principle of abstraction: (PA) Given any condition  $C$  there is a class  $\{x: C(x)\}$  whose members are just those classes  $x$  such that  $C(x)$ . Namely, (2) is a special case of (PA) and (1) follows from (PA) with  $x = x$  as  $C(x)$ . On the other hand, if we substitute  $U$  which is supposed to exist by (1), in place of  $K$  at (2) we get (PA).)

Even before they faced the contradiction, Cantorians refuted (1), because they knew that the class of all classes is an inconsistent multiplicity, the unincreasable absolute infinite. Russell refuted both assumptions, (1) and (2), in his theory of types. Quine refuted (2) in his 'New foundations for mathematical logic' by restricting  $C(x)$  to stratified conditions, but he did not refute (1). The Universe is a legitimate class in Quine, 1937.

## 4

If we apply the diagonal argument to the properties of natural numbers we conclude that there are more properties of natural numbers than natural numbers. Namely, let us suppose that all the properties could be arranged in a sequence  $P_1, P_2, \dots$ , which can be visualized as an infinite rectangular array:

	1	2	3	4	...
$P_1$	<i>t</i>	<i>f</i>	<i>t</i>	<i>f</i>	...
$P_2$	<i>t</i>	<i>f</i>	<i>f</i>	<i>t</i>	...
$P_3$	<i>f</i>	<i>t</i>	<i>t</i>	<i>f</i>	...
$P_4$	<i>t</i>	<i>f</i>	<i>t</i>	<i>t</i>	...
⋮	⋮	⋮	⋮	⋮	⋮
⋮	⋮	⋮	⋮	⋮	⋮

Each row determines those numbers of which the respective property is true. In this particular case  $P_1$  is true of 1,3, ..., and false of 2,4, ... . The diagonal of the array determines the antidiagonal property  $A$  which is true of  $n$  if and only if  $P_n$  is false of  $n$ , i.e. for every  $n$ :

$$A(n) \leftrightarrow \neg P_n(n). \quad (10)$$

From the assumption that all the properties belong to our sequence, it follows that  $P_k = A$ , for some  $k$ . From (10) and  $P_k = A$  the contradiction follows:

$$P_k(k) \leftrightarrow A(k) \leftrightarrow \neg P_k(k). \quad (11)$$

Hence, the properties could not be arranged in a sequence. There are more properties than natural numbers.

But if we think about the properties as expressible in a language, then we are not ready to question their denumerability. Namely, there are only denumerably many expressions (i.e. finite strings of symbols) which we use to express the properties in the language. Nevertheless, the antidiagonal property  $A$  expressed by (10) is different from all the properties arranged in a sequence  $P_1, P_2, \dots$ , of all expressible properties. Is this a contradiction? Richard argued (in a similar situation) that it is not, because the property expressed by (10) presupposes an infinite sequence of properties, and so could be 'expressed' only with infinitely many words (Richard, 1905).

In the late 1920s, when Gödel and Tarski started to consider such questions, it became natural to formulate the questions in the framework of formal systems. It is more precise and more general, in a way, to restrict oneself to a formal system and reckon its one-place formulae as properties of whatever the system is referring to. Gödel taught us that reasonably strong systems of arithmetic may refer to their own formulae via the so-called Gödel numbers, i.e. they contain a name ' $F$ ' for any of its formulae  $F$ . The name ' $F$ ' is the Gödel number of  $F$ , if our formal system is a system of arithmetic, but it could also be simply introduced into the system as it is done in any system of linguistics. Denumerably many one-place formulae of the system could be arranged in a sequence  $F, G, H, \dots$ , and it is legitimate to ask if any of these formulae is true of ' $F$ ', or of ' $G$ ', or of ' $H$ ', etc. The answers can be visualized as an infinite rectangular array:

	'F'	'G'	'H'	...
F	<i>t</i>	<i>f</i>	<i>t</i>	...
G	<i>f</i>	<i>t</i>	<i>f</i>	...
H	<i>t</i>	<i>t</i>	<i>f</i>	...
⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	

If there is a truth predicate *T* in our system, i.e. if there is a one-place formula *T*, such that for every formula *A*

$$T('A') \leftrightarrow A, \tag{12}$$

then our array is of the following form:

	'F'	'G'	'H'	...
F	$T('F('F'))$	$T('F('G'))$	$T('F('H'))$	...
G	$T('G('F'))$	$T('G('G'))$	$T('G('H'))$	...
H	$T('H('F'))$	$T('H('G'))$	$T('H('H'))$	...
⋮	⋮	⋮	⋮	
⋮	⋮	⋮	⋮	

The antidiagonal property *A* acquires the sequence of truth values of the following sentences:

$$\neg T('F('F)'), \quad \neg T('G('G)'), \quad \neg T('H('H)'), \quad \dots ;$$

when tested on the following sequence of arguments:

$$'F', 'G', 'H', \dots$$

A function *d* such that, for any one-place formula *A*,

$$d('A') = 'A('A')', \tag{13}$$

is called the diagonal function. If *d* is expressible in our system then the antidiagonal property *A* is also expressible in our system as  $\neg T(d(x))$ . Hence,  $\neg T(d(x))$  is one of the formulae in our sequence *F*, *G*, *H*, ... . But a contradiction follows if we substitute the corresponding argument into the formula, cf. (11). The corresponding argument of a one-place formula *F* is its name '*F*', so we have to substitute ' $\neg T(d(x))$ ' for *x* in  $\neg T(d(x))$ . If we denote:

$$\neg T(d(' \neg T(d(x)) ')) = A, \tag{14}$$

then it follows from (13) that

$$d(' \neg T(d(x)) ') = 'A'. \tag{15}$$

From (14) and (15) it follows

$$\neg T('A') \leftrightarrow A$$

what contradicts (12). Note that *A* asserts its own falsity, i.e. *A* is the liar sentence.

There are several assumptions in our derivation of the contradiction:

1. The formal system contains the names of its formulae.
2. The diagonal function *d* is expressible in the system.
3. The truth-predicate *T* is expressible in the system.

A reasonably strong system of arithmetic always satisfies (1) and (2), so we have to conclude that the truth-predicate  $T$  is not expressible in such a system. This is the famous Tarski's theorem on the undefinability of truth (Tarski, 1935).

Instead of the undefinable truth-predicate  $T$  Gödel used the definable proveability predicate  $B$  (*beweisbar*). Such a predicate is definable in reasonably strong systems of arithmetic, as Gödel proved (Gödel, 1931). If we denote:

$$\neg B(d(\neg B(d(x)))) = G \quad (16)$$

then it follows from (13) that

$$d(\neg B(d(x))) = 'G'. \quad (17)$$

From (16) and (17) it follows that

$$\neg B('G') \leftrightarrow G. \quad (18)$$

Hence, in a reasonably strong system of arithmetic there is a sentence which asserts its own unproveability. This is the content of the well-known Gödel's diagonal lemma. Gödel also proved (Gödel, 1931):

1. If any sentence  $G$  is proveable in the system then it is proveable that it is proveable
 
$$\vdash G \rightarrow \vdash B('G'). \quad (19)$$

2. If it is proveable in the system that a sentence  $G$  is proveable then it is proveable
 
$$\vdash B('G') \rightarrow \vdash G. \quad (20)$$

From (18) and (19) it follows

$$\vdash G \rightarrow \vdash B('G') \rightarrow \vdash \neg G,$$

and from (18) and (20) it follows

$$\vdash \neg G \rightarrow \vdash B('G') \rightarrow \vdash G.$$

Hence, if the system is consistent (i.e.  $\neg(G \ \& \ \neg G)$ ) then it is incomplete (i.e.  $\nvdash G$  and  $\nvdash \neg G$ ). This is the famous Gödel's incompleteness theorem.

## 5

It is well known that the diagonal argument plays a prominent role in Cantorian set theory and logicist accounts of mathematics, as well as in proving undecidability results in computability theory and incompleteness results in logic and metamathematics. But we hope that the simplicity and the continuity of the interconnections between the areas, which we have tried to reveal by simple and continuous transformations of the diagonal argument, has shed some new light on this whole interplay of ideas.

## Notes

1. The title of (Cantor, 1874) is curious to many historians of set theory, because it emphasizes a property of the collection of all real algebraic numbers (i.e. its denumerability) but not the more important property of real numbers (i.e. its non-denumerability) which is also proved in the paper. The usual explanation is that Cantor tried to minimize those features of his proof that might raise any questions about its acceptability for publication (perhaps by Kronecker). But it is also possible that the title really refers to the most important property of the collection of all real algebraic numbers, its increasability.

2. Cantor also used his number-classes to create a series of increasable infinities. This was in essence the modern ordinal theory of cardinal numbers. The diagonal argument enters this approach through the paradox of the ordinal of all ordinals (cf. note 5).
3. In what follows we use the term class synonymously with set, collection, etc. (If we want to deny the classhood or sethood of a multiplicity we use the term inconsistent class.) We also suppose that all objects we are talking about (in this paragraph) are classes. (If we need individuals we may agree with Quine to treat them as classes  $x$  such that  $x = \{x\}$ .)
4. After they quote Russell's letter to Couturat in which Russell asserts that Cantor's proof presuppose that there are classes included in  $U$  which are not members of  $U$ , Moore and Garciadiego (1981) continue with the following sentence: 'Why Russell believed Cantor's proof to make such a presupposition is not at all clear'. It is not at all clear why they think that it is not clear. Namely, without such a presupposition Zermelo's derivation of the contradiction is immediate.
5. Sometimes, the inconsistency of a class could be quite interesting. If we suppose (as it seems that Cantor did) that all inconsistent, unincreasable, absolutely infinite classes are of the same size, i.e. that there is a one-one correspondence between any two of them, then the inconsistencies of the class of all classes and the class of all ordinals implies that there is a one-one correspondence between them. This proves that the universe of all classes is well-ordered.

## References

- CANTOR, G. (1874) Über eine Eigenschaft des Inbegriffes aller reellen algebraischen Zahlen, *Journal für die Reine und Angewandte Mathematik* 77, pp. 258–262.
- CANTOR, G. (1891) Über eine elementare Frage der Mannigfaltigkeitslehre, *Jahresbericht der Deutschen Mathematiker-Vereinigung* 1, pp. 75–78.
- FREGE, G. (1893) *Grundgesetze der Arithmetik Begriffsschriftlich Abgeleitet* (Verlag, Hermann Pohle).
- GÖDEL, K. (1931) Über formal unentscheidbare Sätze der Principia mathematica und verwandter Systeme I, *Monatshefte für Mathematik und Physik* 38, pp. 173–198 (tr. in VAN HEIJENOORT, 1967).
- HALLET, M. (1984) *Cantorian Set Theory and Limitation of Size* (Oxford, Oxford University Press).
- VAN HEIJENOORT, J. (1967) *From Frege to Gödel* (Harvard, Harvard University Press).
- KLEENE, S. K. (1936) General Recursive Functions of Natural Numbers, *Mathematische Annalen* 112, pp. 727–742.
- MOORE, G. H. & GARCIADIEGO, A. (1981) Buralli-Forti's Paradox: a Repraisal of its Origins, *History of Mathematics* 8, pp. 319–350.
- QUINE, W. V. O. (1937) New Foundations for Mathematical Logic, *Am. Math. Month.* 44, pp. 70–80.
- RANG, B. & THOMAS, W. Zermelo's Discovery of the 'Russell Paradox', *History of Mathematics* 8, pp. 15–22.
- RICE, H. G. (1953) Classes of Recursively Enumerable Sets and Their Decision Problems, *Trans. Am. Math. Soc.* 74, pp. 358–366.
- RICHARD, J. (1905) Les principes des mathématiques et le problème des ensembles, *Revue Générale des Sciences Pures et Appliquées* 16, pp. 541.
- RUSSELL, B. (1903) *Principles of Mathematics* (Cambridge, Cambridge University Press).
- ŠICKIĆ, Z. (1986) Cantor's Theorem and Paradoxical Classes, *Zeitschrift für Mathematische Logik und Grundlagen der Mathematik* 32, pp. 221–226.
- TARSKI, A. (1935) Der Wahrheitsbegriff in den formalisierten Sprachen (tr. in *Logic, Semantics, Metamathematics, Papers from 1928 to 1938* (Oxford, Oxford University Press, 1956).
- TURING, A. M. (1936) On Computable Numbers with an Application to the Entscheidungsproblem, *Proceedings of the London Mathematics Society* 42, pp. 230–265.
- ZERMELO, E. (1908) Neuer Beweis für die Möglichkeit einer Wohlordnung. *Mathematische Annalen* 65, pp. 261–281 (tr. in VAN HEIJENOORT, 1967).