Understanding the Infinite

Shaughan Lavine

Harvard University Press
Cambridge, Massachusetts
London, England

Printed in the United States of America
Second printing, 1998

First Harvard University Press paperback edition, 1998

Library of Congress Cataloging-in-Publication Data
Lavine, Shaughan.
Understanding the infinite / Shaughan Lavine.
p. cm.
Includes bibliographical references and index.
ISBN 0-674-92096-1 (cloth)
ISBN 0-674-92117-8 (pbk.)
1. Mathematics—Philosophy. I. Title.
QA8.4.L38 1994
511.3'22—dc20
93-49697

Preface

In writing this book I have tried to keep mathematical prerequisites to a minimum. The reader who is essentially innocent of mathematical knowledge beyond that taught in high school should be able to read at least halfway through Chapter VIII plus parts of the rest of the book, though such a reader will need to skip the occasional formula. That is enough of the book for all of the major ideas to be presented. The Introduction may seem daunting since it refers to ideas that are not explained until later—trust me, they *are* explained. A reader who learned freshman calculus once, but perhaps does not remember it very well, and who has had a logic course that included a proof of the completeness theorem will be in fine shape throughout the book, except for various "technical remarks," an appendix to Chapter VI, and a few parts of Chapter IX. Those few technical discussions require varying degrees of mathematical sophistication and knowledge of general mathematical logic plus occasional knowledge of elementary recursion theory, model theory, or modal logic.

Thanks are due Bonnie Kent, Vann McGee, Sidney Morgenbesser, and Sarah Stebbins for their infinite patience in listening to my many half-baked ideas and for their substantial help in culling and completing them while I was writing this book. As they learned, I cannot think without the give and take of conversation. Thanks also to Ti-Grace Atkinson, Jeff Barrett, William Boos, Hartry Field, Alan Gabbey, Haim Gaifman, Alexander George, Allen Hazen, Gregory Landini, Penelope Maddy, Robert Miller, Edward Nelson, Ahmet Omurtag, David Owen, Charles Parsons, Thomas Pogge, Vincent Renzi, Scott Shapiro, Mark Steiner, and Robert Vaught for their thoughtful comments on an early version of the book. Those comments have led to significant improvements. And thanks to Thomas Pogge for his substantial help in correcting my translations from German. Any remaining mistakes are, of course, my own.

Thanks are also due my parents, Dorothy and Leroy Lavine, not only for their moral support, which I very much appreciated, but also for their generous financial support, without which the preparation of this book would not have been possible. My wife, Caroline, and daughter, Caila, deserve the most special thanks of all, for tolerating with such understanding my absences and the stresses on our family life that the writing of a book inevitably required. This book is dedicated to them.

Contents

I

Introduction

In the latter half of the nineteenth century Georg Cantor introduced the infinite into mathematics. The Cantorian infinite has been one of the main nutrients for the spectacular flowering of mathematics in the twentieth century, and yet it remains mysterious and ill understood.

At some point during the 1870s Cantor realized that sets—that is, collections in a familiar sense that had always been a part of mathematics—were worthy of study in their own right. He developed a theory of the sizes of infinite collections and an infinite arithmetic to serve as a generalization of ordinary arithmetic. He generalized his theory of sets so that it could encompass all of mathematics. The theory has become crucial for both mathematics and the philosophy of mathematics as a result. Unfortunately, Cantor had been naive, as Cantor himself and Cesare Burali-Forti realized late in the nineteenth century and as Bertrand Russell realized early in the twentieth. His simple and elegant set theory was inconsistent—it was subject to paradoxes.

The history of set theory ever since the discovery of the paradoxes has been one of attempting to salvage as much as possible of Cantor's naive theory. Formal axiom systems have been developed in order to codify a somewhat arbitrarily restricted part of Cantor's simple theory, formal systems that have two virtues: they permit a reconstruction of much of Cantor's positive work, and they are, we hope, consistent. At least the axiomatic theories have been formulated to avoid all of the known pitfalls. Nonetheless, they involve certain undesirable features: First, the Axiom of Choice is a part of the theories not so much because it seems true—it is at best controversial—but because it seems to be required to get the desired results. Second, since present-day set theory is *ad hoc,* the result of retreat from disaster, we cannot expect it to correspond in any very simple way to our uneducated intuitions about collections. Those are what got Cantor into trouble in the first place.

We can never rely on our intuitions again. The fundamental axioms of mathematics—those of the set theory that is its modern basis—are to a large extent arbitrary and historically determined. They are the remote and imperfectly inferred remnants of Cantor's beautiful but tragically flawed paradise.

The story I have just told is a common one, widely believed. Not one word of it is true. That is important, not just for the history of mathematics but for the philosophy of mathematics and many other parts of philosophy as well. The story has influenced our ideas about the mathematical infinite, and hence our ideas about mathematics and about abstract knowledge in general, in many deep ways.

Both elementary number theory and the geometry of the Greeks, for all that they are abstract, have clear ties to experience. They are, in fact, often thought to result from idealizing that experience. Modern mathematics, including much of the mathematics of physics, is frequently thought to be abstract in a much more thoroughgoing sense. As I shall put it, modern mathematics is not only abstract but also remote, because it is set-theoretic:[1] The story tells us that modern axiomatic set theory is the product not of idealization but of the failure of an attempted idealization.

Since science and often mathematics are thought of as quintessential examples of human knowledge, modern epistemology tries to come to grips with scientific and mathematical knowledge, to see it as knowledge of a typical or core kind. That poses a serious problem for epistemology, since mathematical knowledge and the scientific knowledge that incorporates it is thought to be so remote.

The whole picture of mathematical knowledge that drives the epistemology is wrong. As this book will demonstrate, set theory, as Cantor and Ernst Zermelo developed it, is connected to a kind of idealization from human experience much like that connected to the numbers or to Euclidean geometry.

Cantor studied the theory of trigonometric series during the 1870s. He became interested in arbitrary sets of real numbers in the process of making

that theory apply to more general classes of functions. His work was part of a long historical development that had in his day culminated in the idea that a function from the real numbers to the real numbers is just any association—however arbitrary—from each real number to a single other real number, the value of the function. The term *arbitrary* is to make it clear that no rule or method of computation need be involved. That notion of a function is the one we use today.

Cantor's study of the theory of trigonometric series led him to this progression of transfinite "indexes":

$$0, 1, \ldots, \infty, \infty + 1, \infty + 2, \ldots, \infty \cdot 2, \ldots, \infty \cdot 3, \ldots,$$
$$\infty^2, \ldots, \infty^3, \ldots, \infty^\infty, \ldots, \infty^{\infty^\infty}, \ldots.$$

Cantor's set theory began as, and always remained, an attempt to work out the consequences of the progression, especially the consequences for sets of real numbers. Despite the usual story, Cantor's set theory was a theory not of collections in some familiar sense but of collections that can be counted using the indexes—the finite and transfinite ordinal numbers, as he came to call them. Though Cantor came to realize the general utility of his theory for codifying a large part of mathematics, that was never his main goal.

Cantor's original set theory was neither naive nor subject to paradoxes. It grew seamlessly out of a single coherent idea: sets are collections *that can be counted*. He treated infinite collections as if they were finite to such an extent that the most sensitive historian of Cantor's work, Michael Hallett, wrote of Cantor's "finitism." Cantor's theory is a part of the one we use today.

Russell was the inventor of the naive set theory so often attributed to Cantor. Russell was building on work of Giuseppe Peano. Russell was also the one to discover paradoxes in the naive set theory he had invented. Cantor, when he learned of the paradoxes, simply observed that they did not apply to his own theory. He never worried about them, since they had nothing to do with him. Burali-Forti didn't discover any paradoxes either, though his work suggested a paradox to Russell.

Cantor's theory had other problems. It did not, in its original form, include the real numbers as a set. Cantor had, for good reason, believed until the 1890s—very late in his career—that it would include them. (Most everything else I am saying here is known to one or another historian or mathematician, but the claim that Cantor had a smooth theory that broke down in the 1890s is

[1] When I say that modern mathematics is set-theoretic, I am not referring to the so-called set-theoretic foundations of mathematics, which play little role in this book. What I have in mind is the ubiquitous use of set-theoretic concepts in mathematics, concepts like open set, closed set, countable set, abstract structure, and so on and on. The concepts mentioned were, as we shall see in Chapter III, introduced by Cantor in the course of the same investigations in which he introduced his theory of infinite numbers and their arithmetic.

new here. It is argued in detail in §IV.2.)[2] Cantor grafted a new assumption on to his theory as soon as he realized he needed it, an assumption that allowed him to incorporate the real numbers, but the assumption caused big trouble.

The new assumption was his version of what is today the Power Set Axiom. The trouble it caused was that his theory was supposed to be a theory of collections that can be counted, but he did not know how to count the new collections to which the Power Set Axiom gave rise. The whole theory was therefore thrown into doubt, but not, let me emphasize, into contradiction and paradox. It seemed that counting could no longer serve as the key idea. Cantor did not know how to replace it.

Zermelo came to the rescue of Cantor's theory of sets in 1904. He isolated a principle inherent in the notion of an arbitrary function, a principle that had been used without special note by many mathematicians, including Cantor, in the study of functions and that had also been used by Cantor in his study of the ordinal numbers. Zermelo named that principle the Axiom of Choice. Though the principle had been used before Zermelo without special notice, no oversight had been involved: the principle really is inherent in the notion of an arbitrary function. What Zermelo noted was that the principle could be used to "count," in the Cantorian sense, those collections that had given Cantor so much trouble, which restored a certain unity to set theory.

The Axiom of Choice was never, despite the usual story, a source of controversy. Everyone agreed that it is a part of the notion of an arbitrary function. The brouhaha that attended Zermelo's introduction of Choice was a dispute about whether the notion of an arbitrary function was the appropriate one to use in mathematics (and indeed about whether it was a coherent notion). The rival idea was that functions should be taken to be given only by rules, an idea that would put Choice in doubt. The controversy was between advocates of taking mathematics to be about arbitrary functions and advocates of taking mathematics to be about functions given by rules—not about Choice *per se,* but about the correct notion of function. Arbitrary functions have won, and Choice comes with them. There is, therefore, no longer any reason to think of the Axiom of Choice as in any way questionable.

Zermelo's work was widely criticized. One important criticism was that he had used principles that, like Russell's, led to known contradictions. He hadn't. In order to defend his theorem that the real numbers can be "counted,"

[2] The reference is to Chapter IV, Section 2. A reference to §2 would be a reference to Section 2 of the present chapter.

Zermelo gave an axiomatic presentation of set theory and a new proof of the theorem on the basis of his axioms. The axioms were to help make it clear that he had been working on the basis of a straightforwardly consistent picture all along. That is a far cry from the common view that he axiomatized set theory to provide a consistent theory in the absence of any apparent way out of the paradoxes.

There *was* a theory developed as a retreat from the disastrous Russellian theory and its precursor in Gottlob Frege, namely, the theory of types. But it never had much to do with Cantorian set theory. I discuss it only in so far as that is necessary to distinguish it from Cantorian set theory. In the process of discussing it, I introduce a distinctive use Russell suggested for something like schemas,[3] a use that shows that schemas have useful properties deserving of more serious study. Such a study is a running subtheme of this book.

It did not take long for Thoralf Skolem and Abraham Fraenkel to note that Zermelo's axioms, while they served Zermelo's purpose of defending his theorem, were missing an important principle of Cantorian set theory—what is now the Replacement Axiom. The universal agreement about the truth of the Replacement Axiom that followed is remarkable, since the axiom wasn't good for anything. That is, at a time when Replacement was not known to have any consequences about anything except the properties of the higher reaches of the Cantorian infinite, it was nonetheless immediately and universally accepted as a correct principle about Cantorian sets.

Chapters II–V establish in considerable detail that it is the historical sketch just given that is correct, not the usual one I parodied above, and they include other details of the development of set theory. Just one more sample—the iterative conception of set, which is today often taken to be the conception that motivated the development of set theory and to be the one that justifies the axioms, was not so much as suggested, let alone advocated by anyone, until 1947.

There are three main philosophical purposes for telling the story just sketched. The first is to counteract the baneful influence of the standard account, which seems to have convinced many philosophers of mathematics that our intuitions are seriously defective and not to be relied on and that the axioms of mathematics are therefore to a large extent arbitrary, historically

[3] A *schema* is a statement form used to suggest a list of statements. For example, $X = X$, where the substitution class for X is numerals, is a schema that has as instances, among others, $0 = 0$, $1 = 1$, and $2 = 2$.

determined, conventional, and so forth. The details vary, but the pejoratives multiply.

On the contrary, set theory is not riddled with paradoxes. It was never in such dire straits. It developed in a fairly direct way as the unfolding of a more or less coherent conception. (Actually, I think there have been two main strands in the development of the theory, symbolized above by the notion of counting and by Power Set. As I discuss in §V.5, it could be clearer how they fit together. One symptom of our lack of clarity on the issue is the independence of the Continuum Hypothesis. But that is a far cry from the usual tale of woe.)

The second purpose is to show what as a matter of historical fact we know about the Cantorian infinite on the basis of clear and universal intuitions that distinctively concern the infinite. The two most striking cases of things we know about the Cantorian infinite on the basis of intuition are codified as Choice and Replacement. How we could know such things? It seems completely mysterious. The verdict has often been that we do not—our use of Choice and Replacement is to a large extent arbitrary, historically determined, conventional, and so forth. But that is not true to the historical facts of mathematical practice, facts that any adequate philosophy of mathematics must confront. (Allow me to take the liberty of ignoring constructivist skepticism about such matters in the Introduction. I shall confront it in the text.)

The third purpose is to make clearer the nature of intuition—the basis on which we know what we do. I have been using the term *intuition* because it is so familiar, but I do not mean the sort of armchair contemplation of a Platonic heaven or the occult form of perception that the term conjures up for many. Whatever intuition is, it is very important to mathematics:

> In mathematics, as in any scientific research, we find two tendencies present. On the one hand, the tendency toward *abstraction* . . . On the other hand, the tendency toward *intuitive understanding* fosters a more immediate grasp of the objects one studies, a live *rapport* with them, so to speak, which stresses the concrete meaning of their relations.
>
> . . . It is still as true today as it ever was that *intuitive* understanding plays a major role in geometry. And such concrete intuition is of great value not only for the research worker, but also for anyone who wishes to study and appreciate the results of research in geometry. (Page iii of David Hilbert's preface to [HCV52].)

The quotation is from a book about geometry, but the point is far more general.

Just as one scientific theory can displace another because of its superior ability to systematize, one mathematical theory can displace another. Unexpected developments can spawn new theories, which can in turn lead to fruitful developments in old theories and become so intertwined with them that the new and the old become indistinguishable. We shall see examples of those things: The modern notion of a function evolved gradually out of the desire to see what curves can be represented as trigonometric series. The study of arbitrary functions, in the modern sense, led Cantor to the ordinal numbers, which led to set theory. And set theory became so intertwined with the theories of functions and of the real numbers as to transform them completely. That is all a part of the story told in Chapters II and III. Mathematics does not have the same ties to experiment as science, but the way mathematics evolves is nonetheless very similar to the way that science evolves.

The view of mathematics just outlined is usually thought to be antithetical to the possibility of any distinctive sort of mathematical intuition. New mathematics has been thought to evolve out of old without any further constraint than what can be proved. But that cannot have been right for most of the history of modern mathematics: from, say, the first half of the seventeenth century until the second half of the nineteenth there was no coherent systematization or axiomatization for much of mathematics and certainly no adequate notion of proof.

Mathematicians necessarily saw themselves as working on the basis of an intuitive conception, relying to some extent on what was obvious, to some extent on connections with physics, and to some extent—but only to some extent, since proof was not a completely reliable procedure—on proof. (See Chapter II.) I believe that most mathematicians today still see themselves as working in much the same conceptually based and quasi-intuitive way, though that is much harder to show, since rigorous standards of proof and precise axiomatizations are now available. The intuitive conceptions that underlie mathematical theories evolve, as do the theories, but the intuitions both constrain the theories and suggest new developments in them in unexpected ways.

The development of set theory is an excellent example of the positive and necessary role intuition plays in mathematics. Because set theory is in so many respects unlike the mathematics that had gone before, it is clear that prior training was far from an adequate guide for Cantor. Besides, the progression that he found does, in some sense, have clear intuitive content. There is a great and mysterious puzzle in the suggestiveness of Cantor's progression that can hardly be overstated. The progression is infinite, and we have absolutely no experience of any kind of the infinite. So what method are we using—what method did Cantor use—to make sense of the progression? The question is another version of the one raised above about Choice and Replacement.

It is difficult to understand how we can know any mathematical truths at all, since the subject matter of mathematics is so abstract. But the problem is particularly acute for truths about the infinite. There is no doubt that we know that $2+2=4$ in some sense or other, and that that knowledge is somehow connected to our experience that disjoint pairs combine to form a quadruple. The facts are indisputable and have multifarious connections to human experience. But there *is* genuine doubt about the truth of, say, $\aleph_2 + \aleph_2 = \aleph_2$, because, for example, there is doubt about whether there could be $\aleph_2$ things.[4] Everyone agrees we must in some sense accept that $2+2=4$, but it is reasonable to be altogether skeptical about the infinite. Worse still, it is not clear what connections to human experience truths about the infinite might have. A modern philosopher of mathematics put it this way:

> The human mind is finite and the set theoretic hierarchy is infinite. Presumably any contact between my mind and the iterative hierarchy can involve at most finitely much of the latter structure. But in that case, I might just as well be related to any one of a host of other structures that agree with the standard hierarchy only on the minuscule finite portion I've managed to grasp. [Mad90, p. 79]

There is a general philosophical problem about knowledge of abstract objects, mathematical objects in particular. But the special case of knowledge of infinite mathematical objects is a distinctive problem for which distinctive solutions have been suggested. Chapters VI and VII are concerned with that problem of the infinite. In Chapter VI, I survey various accounts of mathematical knowledge of the infinite that attempt to show how it can come out of experience. They begin with a theory of knowledge and try to fit mathematics to it. Intuitionism, various forms of formalism, and one version of David Hilbert's program are discussed. I use a Russellian picture of schemas to clarify how Hilbert's finitary mathematics could avoid any commitment to the infinite. It is a consequence of each of the philosophies surveyed that we could not know what we in fact do.

[4] The symbol is a capital Hebrew aleph. $\aleph_2$ (pronounced "aleph two") stands for one of Cantor's infinite numbers.

In Chapter VII, I survey various accounts of mathematical knowledge of the infinite that go in the opposite direction. They begin with mathematics and try to fit a theory of knowledge to it. Kurt Gödel's views and those of Willard Van Orman Quine and Hilary Putnam are discussed. Each fails to account for the higher reaches of set theory. I also discuss Skolem's skeptical challenge to mathematical knowledge of the infinite—a history of which is a part of Chapter V—and the attempt to use second-order logic to block it. While I conclude that the Skolemite criticism of second-order logic has merit, I propose a related solution to the skeptical problem, one dependent on the use of schemas, that I believe succeeds.

None of the philosophies discussed in Chapters VI and VII could solve the problem of the infinite because none of them faced up to the main issue—What is the source of our intuitions concerning the Cantorian infinite? In more prosaic and somewhat over-simple terms, what do the ellipses, the triples of dots, in the written form of Cantor's transfinite progression suggest to us? Whatever that is is a large part of what led Cantor to his theory.

Finding an answer is important for many reasons. Our set theory is incomplete—it is inadequate for resolving many of the problems to which it gives rise. Anything that helps to clarify the sources of our axioms may help to suggest more axioms or help to adjudicate between the additional ones that have already been proposed. That is important both for mathematical reasons and because the apparent hopelessness of finding new axioms has itself become a source of skepticism about the mathematical theory of the infinite.

The apparent problem in accounting for the mathematical infinite led to the split between the philosophers discussed in Chapter VI and those discussed in Chapter VII. Each side seems today to be a council of despair. The resulting impasse has had repercussions far beyond the philosophy of mathematics. It has affected all modern epistemological theories.

In Chapter VIII, I propose that the source of our intuitions concerning the Cantorian infinite is experience of the indefinitely large. That is, our image of what the ellipses represent arises from our idea of going on for much longer than we have so far—going on indefinitely long. The proposal may gain some plausibility from the fact that children go through a stage at which they think the infinite literally is nothing more than the indefinitely large.

The proposal is nothing new, but I give a substantial new argument for it, making use of a mathematical theory of the indefinitely large developed by Jan Mycielski. In order to show that the theory can serve as a codification of the actual historical and psychological source of our intuitions concerning the infinite, it is necessary to show four things: (1) that the theory does not presuppose the infinite and is therefore suited in principle to be a source of intuitions concerning the infinite in that it does not presuppose what it is to explain; (2) that the theory formalizes ordinary experience of the indefinitely large and is therefore a reconstruction of intuitions that we have, as a matter of actual psychological fact; (3) that it does lead to set theory, and that it is therefore rich enough to explain what we have set out to explain; and (4) that it coheres well with the actual development of set theory, and thus that it can be taken to capture the intuitions that played an actual historical role.

To show the first, that the infinite is not presupposed, it is necessary to present the theory in such a way that it involves no commitment to the infinite. That is done using schemas. As a bonus this presentation shows, using mathematical work of Mycielski, that the theory enables us to provide a counterpart for ordinary set-theoretic mathematics that involves no commitment to the infinite.

To argue for the second, that the theory is a reasonable codification of our experience of the indefinitely large, I show how it can be applied to make some parts of the calculus more obvious—connected with daily experience—than they are when given the usual presentation involving limits. That—in addition to the plausibility of the theory in itself—shows how natural and intuitive the theory is, and, as you will see for yourself, how close to your pre-theoretic intuitions.

I show the third, that the theory does lead to set theory, by showing that set theory, including Choice and Replacement, arises by extrapolation, in a precise mathematical sense, from the theory of the indefinitely large.

The chief argument for the fourth, that the theory coheres well with the actual development of set theory, is that the theory of the indefinitely large helps us to make sense of Cantor's "finitism." Cantor saw himself as making an analogy between the finite and the infinite. We can now make precise sense of that: his procedure, analyzed and reconstructed, was that of extrapolating from the indefinitely large to the infinitely large.

The process of idealization that connects the finite to the infinite will be shown not to be very different in principle from the one that connects pencil dots to geometrical points. Points are, more or less, idealized dots, while infinite sets are, more or less, idealized indefinitely large collections. Thus, set theory is of a piece with arithmetic and geometry: all three have a close association with familiar types of experience. The apparently mysterious character of knowledge of the infinite is dissolved.

II

Infinity, Mathematics' Persistent Suitor

. . . But, from the very nature of an irrational number, it would seem to be necessary to understand the mathematical infinite thoroughly before an adequate theory of irrationals is possible. The appeal to infinite classes is obvious in Dedekind's definition of a cut. Such classes lead to serious logical difficulties.

It depends upon the individual mathematician's level of sophistication whether he regards these difficulties as relevant or of no consequence for the consistent development of mathematics. The courageous analyst goes boldly ahead, piling one Babel on top of another and trusting that no outraged god of reason will confound him and all his works, while the critical logician, peering cynically at the foundations of his brother's imposing skyscraper, makes a rapid mental calculation predicting the date of collapse. In the meantime all are busy and all seem to be enjoying themselves. But one conclusion appears to be inescapable: without a consistent theory of the mathematical infinite there is no theory of irrationals; without a theory of irrationals there is no mathematical analysis in any form even remotely resembling what we now have; and finally, without analysis the major part of mathematics—including geometry and most of applied mathematics—as it now exists would cease to exist.

The most important task confronting mathematicians would therefore seem to be the construction of a satisfactory theory of the infinite . . . If the reader will glance back at Eudoxus' definition of "same ratio" . . . he will see that "infinite difficulties" occur there too . . . Nevertheless some progress has been made since Eudoxus wrote; we are at least beginning to understand the nature of our difficulties.

[Bel37, pp. 521–522]

With this chapter, I hope to make better known a few aspects of the history of the mathematical infinite that are known at least in outline to many mathematicians. The chapter is a work of exposition, not of scholarship. Little of what I shall say is controversial.[1] If I succeed in making the story accessible without introducing detailed knowledge of Fourier series or of the distinction between convergence and uniform convergence, the chapter will have served its purpose.

The modern-day theory of the infinite did not begin with an effort to produce a theory of the infinite, and it did not build on a long history of attempts at mathematical theories of the infinite. It began instead with an attempt to clarify the foundations of analysis and specifically of the calculus—that is, it grew out of the development of our theory of rates of change and of areas under curves. The infinite has entered present-day mathematics in large part as the result of attempts to make sense of the notion of an arbitrary curve or function.

The story of the hugely successful application of analysis to physics is one that is too well known to bear retelling here. Let me simply note that analysis could not in Newton's time and cannot today be regarded as just one among many branches of mathematics: it is the one whose application, especially to physics, has been the most fruitful. It is therefore the branch of mathematics through which mathematics makes its most intimate contact with physics, the sciences, and the natural world.

§1. Incommensurable Lengths, Irrational Numbers

Most of us have been taught at one time or another that Pythagoras discovered that the square root of two is irrational. That is very likely not true, though our historical information concerning the Pythagoreans is sparse. First of all, many of the discoveries of the Pythagoreans are attributed to Pythagoras himself, and it is very likely that some other member of the Pythagorean school made the discovery. Indeed, the discovery is attributed to Hippasus of Metapontium (fifth century B.C.E.) among others. Legend has it that he made the discovery while at sea with the other Pythagoreans and that he was tossed overboard for his trouble. (See [Hea81, vol. 1, pp. 154–157] and [Hea56, vol. 1, pp. 411–414].)

Second, and much more important, the only numbers the Pythagoreans had anything to do with were whole numbers—no rational numbers, and certainly no irrational ones. They knew many things about geometrical proportions between *geometrical* magnitudes. For example, they knew that two strings of the same type and tension whose lengths were in the ratio of three to two would, when plucked, produce notes a musical interval of a fifth apart. The ratio of three to two meant approximately that the two lengths could be measured by a common unit so that one was three times the length of that unit, while the other was twice that length. That was in no way associated with the fractions or rational numbers $3/2$ or $2/3$.

The lengths of the two strings in our example were *commensurable*—measurable by whole-number multiples of a common unit. What the Pythagoreans had discovered was not that the square root of two is irrational but that the side and the diagonal of a square are not commensurable. That made it impossible to continue the Pythagorean program of identifying geometry with the theory of the numbers, which were, for the Greeks, just the whole numbers.

Sometime in the century following the work of Hippasus of Metapontium, Eudoxus gave an ingenious theory of incommensurable ratios, a theory that remains the basis of our understanding today. Incommensurable ratios arose within geometry, and his theory was entirely geometric. Indeed, Eudoxus contrasted geometric magnitudes with numbers, which increase a unit at a time. The main idea of his theory of incommensurable ratios is more or less this: a is in the same ratio to b that c is to d if for any whole numbers n and m, na is less than, equal to, or greater than mb if and only if nc is, respectively, less than, equal to, or greater than md.

Less than a century later, the Eudoxian theory was codified in Book V of Euclid's *Elements*. Book II showed how to do what algebra there was geometrically: Numbers are represented or, probably more accurately, replaced by lengths, angles, areas, and volumes. The product of two lengths is an area; the product of three, a volume. One can add and subtract lengths from lengths, areas from areas, and so forth. Numbers and algebra have in effect been eliminated in favor of geometry, and the foundations of the geometrical theory of ratios or proportions are those of Eudoxus.

The ratios of magnitudes, commensurable and incommensurable, are not

[1] I have relied heavily on Morris Kline's *Mathematical Thought from Ancient to Modern Times* [Kli72] and the articles in *From the Calculus to Set Theory* [GG80b], edited by I. Grattan-Guiness. My analysis of the development of the calculus has been heavily influenced by Philip Kitcher's *The Nature of Mathematical Knowledge* [Kit83]. I have also made some use of Florian Cajori's *History of Mathematics* [Caj85] and Dirk J. Struik's *A Concise History of Mathematics* [Str87]. Various more specialized historical works, cited in the text when necessary, have served as useful correctives.

stand-ins for numbers, rational and irrational. No procedure is given, for example, for adding or multiplying ratios of magnitudes.

Neither are the magnitudes themselves—lengths and the like—stand-ins for rational and irrational numbers. One can add them, but the product of lengths, for example, is an area. Euclid was careful to state (Definition 3) that a ratio can only relate magnitudes of the same kind. That is, in particular, one cannot relate lengths and areas in a ratio. Unlike the product of numbers, a product of lengths is an entity of a different kind.

In Book X, Euclid investigated and classified ratios between lines that we would represent as having lengths of the form $\sqrt{\sqrt{a} \pm \sqrt{b}}$ for commensurable a and b. Ratios between lines that cannot be expressed in that form were not discussed in the *Elements*.

Leonardo of Pisa (Fibonacci) was educated in Africa, and he traveled widely. He reintroduced Euclid's *Elements* and other Greek mathematical works to Europe. He also disseminated Arabic numerals and methods of calculation. In 1220, Leonardo published his discovery that the roots of $x^3 + 2x^2 + 10x = 20$ are not expressible in the form $\sqrt{\sqrt{a} \pm \sqrt{b}}$. The Arabs worked freely with irrational numbers, and Leonardo's discovery showed that not every number could be constructed within the Euclidean strictures of compass and straightedge. But no adequate foundation had been provided for the use of irrational numbers.

In succeeding centuries the use of irrational numbers became increasingly common among European mathematicians, but it was not clear in what sense they were numbers. In his *Arithmetica Integra* (1544) Michael Stifel wrote,

> Since, in proving geometrical figures, . . . irrational numbers . . . prove exactly those things which rational numbers could not prove . . . we are moved and compelled to assert that they truly are numbers . . . On the other hand, other considerations compel us to deny that irrational numbers are numbers at all. To wit, when we seek [to give them a decimal representation] . . . we find that they flee away perpetually, so that not one of them can be apprehended precisely in itself . . . Now that cannot be called a true number which is of such a nature that it lacks precision . . . Therefore, just as an infinite number is not a number, so an irrational number is not a true number, but lies hidden in a kind of cloud of infinity. [Kli72, p. 251]

As we shall see, Stifel's remarks were prescient: the basis of the irrational numbers was not adequately clarified until *infinite* numbers were allowed into mathematics.

The ties to geometry remained strong. Stifel said that "going beyond the cube just as if there were more than three dimensions . . . is against nature" [Kli72, p. 279]. René Descartes, around 1628 (in *Regulae ad Directionem Ingenii*), explicitly allowed irrational numbers for continuous magnitudes. In 1637 Descartes took the product of lengths to be a length, not an area, and viewed polynomials as determining curves [Des54]. (See also [Gro80] and [Mah73].) Newton introduced number as "the abstracted ratio of any quantity, to another quantity of the same kind," including incommensurable ratios, and introduced multiplication, division, and roots in terms of ratios in his university lectures, published in 1707 as *Arithmetica universalis sive de compositione et resolutione arithmetica liber* [Whi67, vol. 2, p. 7].

Until now we have been considering the geometry of straight lines (and rectangles, and so forth) and their magnitudes. We shall now turn to the geometry of curves and the areas they bound. Once more, Eudoxus did basic work that Euclid incorporated in the *Elements*, in Book XII. Archimedes went even further in developing what is called the method of exhaustion. The method remained the only fully worked out and thoroughly justified one for computing areas and volumes until the nineteenth century, but the details are not central to our story.

§2. Newton and Leibniz

In the first half of the seventeenth century various curves were introduced or described by means of motion. That was not new, but this method of description came to play an increasingly central role. In 1615 Marin Mersenne defined the cycloid as the path traced out by a point on the edge of a rolling circle. The cycloid was not new; the definition was. Galileo Galilei showed in *Discorsi e dimostrazione matematiche intorno a due nuove scienze* (1638) that the path of a cannonball was a parabola, and he regarded the curve as the path of a moving point.

Many techniques were devised for computing various properties of curves, in part building on the method of exhaustion: techniques for computing maxima and minima, locating tangent lines, and computing areas and volumes. The mathematicians involved included Pierre Fermat, Descartes, Isaac Barrow, Johann Kepler, Bonaventura Cavalieri, Gilles Personne de Roberval, Evangelista Torricelli, Blaise Pascal, John Wallis, Sir Christopher Wren, William Neile, Gregory of St. Vincent, James Gregory, and Christiaan Huygens. But Isaac Newton and Gottfried Wilhelm Leibniz soon systematized the techniques into the calculus, and so we shall only briefly look at the work of the others.

The new study of curves and motion led to a new definition of the line tangent to a curve (Roberval, *Brieves Observation sur la composition des mouvemens et sur le moyen de trouver les Touchantes des Ligne Courbes*, ca. 1636, published 1693). The Greek definition of a line tangent to a curve is a line touching the curve at a point. Roberval defined a tangent to a curve as the direction of the velocity of a moving point tracing the curve.

In his *Arithmetica Infinitorum* (1655), Wallis studied infinite sums and products. He also gave a correct general definition of the limit of an infinite sequence of numbers, a definition that did not surface again until around 1820. (For example, the limit of the sequence $1, \frac{1}{2}, \frac{1}{4}, \ldots$ is 0. See §5.) Newton studied the *Arithmetica Infinitorum* and used its techniques to convince himself that the binomial theorem—which gives the coefficients of the expansion of $(a+b)^n$ for arbitrary n—also held when n was negative or fractional. In those cases, there are infinitely many coefficients—one obtains an expansion of $(a+b)^{m/n}$ as an infinite sum or *series*. (As an example of a series—though not one derived from the binomial theorem—the limit of the series $1 + \frac{1}{2} + \frac{1}{4} + \ldots$ is 2.) Such series were crucial for Newton's development of the calculus, to which we now turn.

In *De Analysi per Aequationes Numero Terminorum Infinitas* (circulated in 1669, published 1711), Newton gave a considerably more general version of the following derivation: Suppose that the area z under a curve is given by $z = x^2$. (See Figure 1, which is not drawn to scale.) Suppose x increases by a "moment" o, that is, in our present-day Leibnizian terminology, by an infinitesimal.[2] (The term *moment* was presumably suggested by thinking of x as time.) Then the area under the curve increases by ov, and so $z + ov = (x+o)^2$, where the right-hand side is obtained by using $z = x^2$, which we have assumed true, at the point at which the x coordinate has value $(x+o)$. Multiplying out, $z + ov = x^2 + 2ox + o^2$, and since $z = x^2$, it follows that $ov = 2ox + o^2$. We now divide through by o to obtain $v = 2x + o$. At this point, Newton took o "infinitely small" to obtain $y = 2x$, since (from the figure) v is equal to y when o is infinitely small.

[2] The history of analysis from this point on depends heavily on present-day ideas about infinitesimals, on which see §VIII.3. Those ideas are used to adjudicate what arguments have a reasonable reconstruction in modern terms, and hence are to be viewed as correct, and which do not.

As Newton himself admitted, the method is "shortly explained rather than accurately demonstrated." The derivation accomplishes two things at once: First, it shows that the rate of change of x^2 is $2x$ (on the right-hand side we computed the change $(x+o)^2 - x^2$ divided by the "time" o in which the change occurs to obtain the rate of change). Second, it shows that the rate of change of the area z is the curve y bounding that area (on the left-hand side we computed the rate of change of z and obtained y). The equation $y = 2x$ thus asserts that the rate of change $(2x)$ of the area $(z = x^2)$ bounded by a curve (y) is the curve itself. That is Newton's version of the fundamental theorem of the calculus[3]—for $z = x^2$. Newton did not use that example. He made

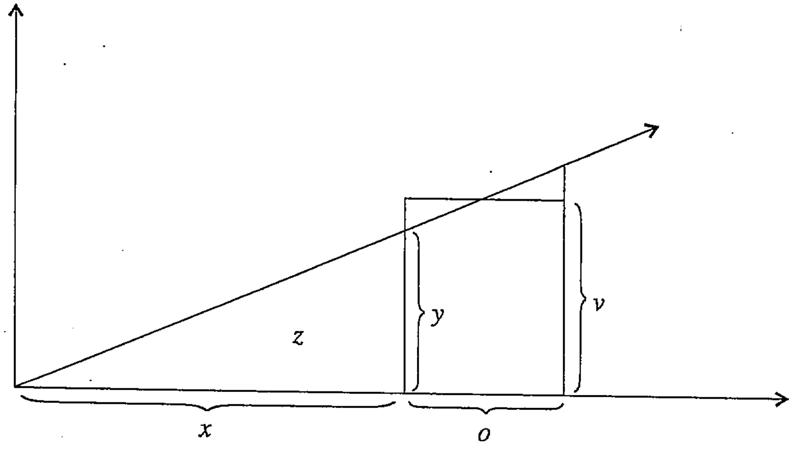

Figure 1. Newton's derivation.

[3] Here is all you need to know about the fundamental theorem of the calculus. I have omitted how to handle negative values since the details don't matter for our story. I have also omitted important restrictions on the applicability of the theorem. They were far from worked out in the days of Newton and Leibniz. Differentiation is pretty much the operation that takes a function f to the function g that graphs the slope or, equivalently, the rate of change of f (that is, $g(x)$ is the slope of f at x or the rate of change of f at x if we think of x as representing time). Integration is pretty much the operation that takes a function f to the function g that graphs the area under f (that is, $g(x)$ is the area under the graph of f between 0 and x). The fundamental theorem of the calculus states that integration and

it clear that one could use $z = ax^m$, where m could be negative or fractional, expanding the right-hand side not by multiplying it out but by using the binomial theorem. He thus obtained the result that the rate of change of ax^m is max^{m-1}. He then expanded other equations involving x as infinite series of terms of the form ax^m and applied the result term by term to compute other rates of change.

In a subsequent work (*Methodus Fluxionum et Serierum Infinitarum,* written in 1671, published 1736), Newton called a variable quantity a "fluent" and its rate of change a "fluxion." He computed rates of change by computing the fluxion of a fluent, and he found areas by finding the fluent of a fluxion. He now regarded fluents as generated by continuous motions instead of as being built up as static assemblages of moments. The moment o is now conveniently thought of as "an infinitely small interval of time." The idea of taking a curve to be the path of a moving point thereby became fundamental. Newton had introduced an early form of the idea of functional dependence—with time as an auxiliary independent variable.

In a third paper (*Tractatus de Quadratura,* written 1676, published 1704), Newton attempted to eliminate the moments, or infinitesimals. He said, "Lines are described . . . not by the apposition of parts, but by the continued motion of points," and "Fluxions are, as near as we please, as the increments of fluents generated in times, equal and as small as possible, and to speak accurately, they are in the prime ratio of nascent increments." His computations were much as before, but the new excuse for dropping terms involving o at the end was "Let now the increments vanish and their last proportion will be . . . " To the modern ear, that phrase suggests the beginnings of the theory of limits that eventually became a crucial part of the foundations of analysis. In contrast to his concern about the increment o, Newton did little to provide a basis for his use of series—infinite sums [Kit83, p. 234].

Let us now turn to Leibniz's independent discovery of the calculus. Whereas Newton relied heavily on temporal ideas and infinite series, Leibniz assimilated curves to sequences of numbers. In 1666, while Newton was completing the main part of his development of the calculus, Leibniz published a work, *De Arte Combinatoria,* on what seems a different subject. Consider the

differentiation are inverse operations, which means that if g is the integral of f, then f is the derivative of g.

following sequences of numbers. Each sequence on a line below another sequence consists of the differences between the terms in the sequence above it:

$$\begin{array}{ccccccccc}
0, & & 1, & & 2, & & 3, & & \dots \\
 & 1, & & 1, & & 1, & & \dots & \\
 & & 0, & & 0, & & \dots & &
\end{array}$$

Now, we start with squares:

$$\begin{array}{ccccccccccc}
0, & & 1, & & 4, & & 9, & & 16, & & \dots \\
 & 1, & & 3, & & 5, & & 7, & & \dots & \\
 & & 2, & & 2, & & 2, & & \dots & & \\
 & & & 0, & & 0 & & \dots & & &
\end{array}$$

Finally, with cubes:

$$\begin{array}{ccccccccccccc}
0, & & 1, & & 8, & & 27, & & 64, & & 125, & & \dots \\
 & 1, & & 7, & & 19, & & 37, & & 61, & & \dots & \\
 & & 6, & & 12, & & 18, & & 24, & & \dots & & \\
 & & & 6, & & 6, & & 6, & & \dots & & & \\
 & & & & 0, & & 0, & & \dots & & & &
\end{array}$$

Leibniz noted that the second differences for the sequence of natural numbers, the third differences for the sequence of squares, and so forth, all vanish. He also recognized that each sequence could be recovered as the successive sums of its first member and the members of the sequence below it—that is, by putting the differences back together. In 1673, during the time between Newton's second paper and his third, Leibniz connected those facts to the study of curves by thinking of a curve as a sequence of successive points. He later came to think of the successive points as differing by infinitesimals. When the succession of points is such that their x coordinates differ by a constant amount, the successive, infinitesimally close x values are thought of as giving the order of the terms in the sequence, while the y values may constitute the terms themselves. Thus, a curve is conceived of in terms of a sequence of values much like the sequences Leibniz had investigated earlier. At this stage, dx (a notation Leibniz introduced a couple of years later) is 1 since the terms are the first, second, third, and so forth, while dy is the actual difference between adjacent terms. He thus saw that if the unit is infinitely small, then the sum of the ys gives the area under the curve and the differences dy ($dy = dy/dx$, since $dx = 1$) are the slopes of the tangent lines. He recognized that (in a now familiar notation he introduced later) $\int dy = y$: the sum of the differences is the original series. That is the beginning of his version of the fundamental theorem of the calculus—the integral ($\int$) of the differential (dy) of y is y. The integral is Leibniz's—and our—term for pretty much what Newton had called a fluent, and the differential played pretty much the role of Newton's fluxion.

Leibniz also made use of the "characteristic triangle," which he adopted from Pascal. (See Figure 2.) The characteristic triangle is abc, where ac is simultaneously a straight line and part of the curve. The curve was in effect viewed as a polygon with infinitesimal sides. The triangle abc is similar to the triangle ABa, which has sides of ordinary finite length. The line Aa is tangent to the curve. Those facts exemplify the main reasons why the characteristic triangle was useful.

Using the above ideas, Leibniz had most of the essential features of his calculus by 1675. The details took a couple of years more. Unlike Newton, Leibniz preferred to avoid the use of infinite series.

At first Leibniz had little to say about the nature of the dxs and dys. In 1680 he said that "these dx and dy are taken to be infinitely small, or the two points on the curve are understood to be a distance apart that is less than any given length." The differential dy is a "momentaneous increment." In 1684, Leibniz defined a tangent as a line joining two points that are infinitely close. In 1690 he said (in a letter to Wallis):

> It is useful to consider quantities infinitely small such that when their ratio is sought, they may not be considered zero but which are rejected as often as they occur with quantities incomparably greater. Thus if we have $x + dx$, dx is rejected. But it is different if we seek the difference between $x + dx$ and x. [Kli72, p. 385]

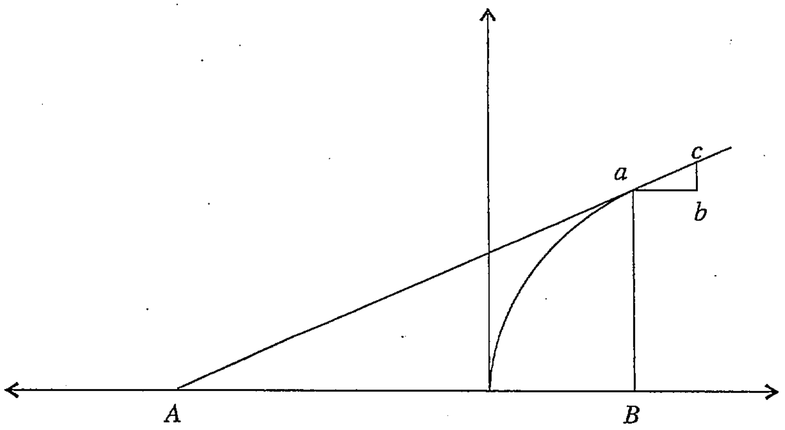

Figure 2. The characteristic triangle.

The infinitesimals are on occasion taken to be vanishing or incipient quantities, or indefinitely small quantities smaller than any finite quantity.

While Newton and Leibniz were struggling with infinitely small and nascent quantities, Wallis was fairly clear about the nature of the number line. He accepted irrationals as numbers, and he thought of the Eudoxian theory of ratios of magnitudes, as presented in Book V of Euclid's *Elements*, as arithmetical. He identified rational numbers with repeating decimals. But the calculus became such a central part of mathematics that the unclarity of its basic concepts infected virtually all of the work of mathematicians. Proof was almost completely abandoned.

In 1673 Leibniz had taken a curve to be given by an equation, but he called any quantity varying along the curve—for example, the length of the tangent line from the curve to the x axis—a function. The function is not, however, a function of a variable but of the curve [Bos74, p. 9]. Newton, at least in principle, did not give a curve any special status: he took fluxions of fluents and fluents of fluxions equally. What Newton considered were quantities obtained from others by means of (possibly infinite) algebraic combinations, primarily infinite sums of finite combinations—what today would be called series.

When one considers infinite series it is necessary to consider convergence if one wishes to avoid absurd results. For example, the series

$$x + x^2 + x^3 + \cdots$$

diverges when $x = 2$, and so does not have a value, but for $x = \frac{1}{2}$, it becomes

$$\frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \cdots,$$

which converges to 1. The terms *convergent* and *divergent* appeared in the work of James Gregory while Newton was developing the calculus. In that period, Lord Brouncker showed some series to be convergent.

Newton did not show as much facility with the distinction between convergent and divergent series as Gregory and Brouncker. He noted that some series (like the one of our example) should be used only for small values of x, while others should be used only for large values of x. He noted that some series become infinite at some values of x and that they are useless for those values.

In 1713 Leibniz devised a test for the convergence of some series, but by and large Newton and Leibniz and their successors simply treated series as on a par with finite sums. Indeed, at around 1800 Joseph-Louis Lagrange tried to provide an algebraic basis for analysis that used infinite series without regard for whether or not they converged.

Leibniz and Newton showed at least occasional concern for the convergence of series. In contrast, the Bernoulli brothers, who studied Leibniz's work and corresponded with him often, did extensive work concerning series, and they showed almost no awareness of any need for caution. Wrong results were described as paradoxes. The Bernoullis also made substantial positive contributions, but those are irrelevant to our concerns here, with a few exceptions. From 1697 on, Johann Bernoulli employed the notion of any "quantity" formed from a variable and constants using algebraic and transcendental expressions. He called such a quantity a "function" beginning in 1698, adopting the term that Leibniz had used earlier. I shall refer to such functions as "analytic expressions," a term used by Leonhard Euler [Bos74, p. 10], to emphasize the difference between such expressions and functions in the modern sense.[4] Johann's work marked the beginning of a transition from a focus on the study of geometrical curves to a focus on the study of analytic expressions. Johann also deemphasized the geometrical basis of the notion of an integral as an area by simply defining the integral to be the inverse of the differential. The fundamental theorem of the calculus was thus absorbed into the definition. That style of defining the integral was dominant into the nineteenth century. Euler solidified the change away from geometry in a series of influential textbooks published from 1748 to 1770. He adopted and generalized Johann Bernoulli's definition of a function. (See [Bos80, pp. 73–79].) Johann also did some work on the vibrating-string problem, a problem that we shall be discussing in detail later.

§3. Go Forward, and Faith Will Come to You

Euler investigated infinite sums in the 1730s. I shall give two examples he used to illustrate the difficulties that such series cause. Formal long division

[4] Analytic expressions bear only the most remote of relations to "analytic functions" in the modern technical sense, which are certainly not what I mean.

of polynomials—the process you were taught in high school—yields

$$\frac{1}{(1+x)^2} = 1 - 2x + 3x^2 - 4x^3 + \cdots$$

and

$$\frac{1}{1-x} = 1 + x + x^2 + x^3 + \cdots.$$

Plugging $x = -1$ into the first series, one obtains (because $1/0$ is ∞)[5]

$$\infty = 1 + 2 + 3 + 4 + \cdots.$$

Plugging $x = 2$ into the second series, one obtains

$$-1 = 1 + 2 + 4 + 8 + \cdots.$$

The series for -1 is term by term greater than or equal to the series just above it for ∞, and it is term by term greater after the first two terms. Hence, Euler noted, one might conclude that $-1 > \infty$ and infer that ∞ is in some sense between the positive and negative numbers. He also plugged $x = -1$ into the series for $1/(1-x)$ to obtain

$$\frac{1}{2} = 1 - 1 + 1 - 1 + \cdots,$$

as Leibniz had done. He considered and rejected the idea that one should restrict attention to the sums of convergent series [Kit83, pp. 242–244].

Euler was no idiot. He was very interested in computing the sums of infinite series, and he developed many techniques for computing them that involved divergent series. The problematic series did not get him into trouble because he could always check his results, at least approximately, by summing a few terms of whatever series he was considering. The divergent series that could cause problems also led to too many successes to simply be

[5] Dividing by 0, as in the equation under discussion, can lead to trouble: Since $0 \cdot 0 = 0 \cdot 1$, one might conclude that $0 = 1$. But using ∞ as a notation for $n/0$, for $n > 0$, is not problematic. The mistake here lies in the use of divergent series, not in the division by 0.

dropped [Kit83, p. 250], though he did things with them that later mathematicians would view with horror [Kit83, p. 323n].

Besides his work on series, Euler solidified the separation of analysis from geometry, introduced functions (analytic expressions) of more than one variable, and gave differential coefficients, essentially derivatives, a crucial role. In 1734 Euler considered a notion of function considerably broader than the analytic expressions we have seen before: he allowed a function to be formed by putting together parts of curves, and even allowed curves freely drawn. He also introduced the now familiar notation $f(x)$ for a function of x.

In that same year George Berkeley, the Anglican bishop of Cloyne, published *The Analyst*, a devastating critique of the foundations of analysis.[6] Like Newton, Berkeley had doubts about matters related to infinitesimals, not infinite series. His criticisms in large part sound exactly right to the modern ear. He understood the value of the methods: "The Method of Fluxions is the general key by help whereof the modern mathematicians unlock the secrets of Geometry, and consequently of Nature" [Ber34, p. 66]. Nonetheless, he said that the mathematicians of his age took more pains to apply their principles than to understand them [Ber34, p. 99]. He pointed out that derivations like the one by Newton described above are incoherent, since one divides through by o and later assumes o equal to zero [Ber34, p. 72]. That criticism may be a bit unfair to Newton, who can, as we have seen, be read as having some idea of using something like limits to replace the procedure of setting o equal to zero [Kit83, p. 239n].

Berkeley criticized Newton's theory of fluxions as ultimate ratios of evanescent increments in a familiar passage:

> And what are these fluxions? The velocities of evanescent increments? And what are these evanescent increments? They are neither finite quantities, nor quantities infinitely small, nor yet nothing. May we not call them the ghosts of departed quantities? [Ber34, p. 88]

Leibniz's infinitesimals fared no better:

> [Our modern analysts] consider quantities infinitely less than the least discernible quantity; and others infinitely less than those infinitely small ones; and still others infinitely less than the preceding infinitesimals, and so on without end or limit. [Ber34, p. 68]

[6] Bernard Nieuwentijdt made similar criticisms of the calculus forty years earlier. He was widely read and provoked a reply from Leibniz. See §VIII.3 for a discussion. See [Man89] for information on other early criticisms of the calculus.

He continued:

> Nothing is easier than to devise expressions or notations, for fluxions and infinitesimals of the first, second, third, fourth, and subsequent orders . . . But if we remove the veil and look underneath, if, laying aside the expressions, we set ourselves attentively to consider the things themselves which are supposed to be expressed or marked thereby, we shall discover much emptiness, darkness, and confusion; nay, if I mistake not, direct impossibilities and contradictions. [Ber34, p. 69]

And finally,

> In all this the ultimate drift of the author [Newton] is very clear, but his principles are obscure. [Ber34, p. 94]

In *A Defence of Free-thinking in Mathematics* (1735, a reply to a reply to *The Analyst*), Berkeley summarized the various contemporary attempts at foundations of analysis:

> Some fly to proportions between nothings. Some reject quantities because infinitesimal. Others allow only finite quantities and reject them because inconsiderable. Others place the method of fluxions on a foot with that of *exhaustions*, and admit nothing new therein. Some maintain the clear conception of fluxions. Others hold they can demonstrate about things incomprehensible. Some would prove the algorithm of fluxions by *reductio ad absurdum;* others *a priori*. Some hold the evanescent increments to be real quantities, some to be nothings, some to be limits. As many men, so many minds . . . Some insist the conclusions are true, and therefore the principles . . . Lastly several . . . frankly owned the objections to be unanswerable. [Ber35a, p. 133]

That seems a pretty fair summary of the state of affairs. Euler, by the way, was one who endorsed "proportions between nothings":

> There is no doubt that every quantity can be diminished to such an extent that it vanishes completely and disappears. But an infinitely small quantity is nothing other than a vanishing quantity and therefore the thing itself equals 0. (*Institutiones*, 1755; see [Kli72, p. 429].)

He then went on to explain how dy/dx, which was $0/0$, could have a definite value. Such was the state of the art. The above quote is from one of Euler's textbooks, which overall had a tremendous positive influence in organizing analysis into a coherent study of analytical expressions [Bos80, pp. 53, 76].

The complete lack of rigor came to be considered a virtue:

> But the tables have turned. All reasoning concerned with what common sense knows in advance, serves only to conceal the truth and to weary the reader and is today disregarded. (Alexis Claude Clairaut, *Eléments de géométrie*, 1741; see [Kli72, p. 619].)

The attitude of Clairaut was widespread. Nonetheless, there were a number of attempts to give a more adequate development of the calculus. Those attempts were motivated by particular mathematical problems—not by any desire to prove the obvious more carefully. Euler's texts emphasized "differential coefficients," which are the derivatives that ultimately replaced differentials in the foundations of analysis [Bos80, p. 74]. Jean Le Rond d'Alembert and Benjamin Robins, like Wallis about a century earlier, emphasized limits [Bos80, p. 91]. D'Alembert said, "The theory of limits is the true metaphysics of the calculus." But he never actually worked out a presentation of the calculus on that basis; indeed it is likely that he could not have done so, since he, like Robins, considered the limits of *variables* (variable quantities) not of *functions* with a specified independent variable [Bos80, p. 92]. In the absence of a correct presentation, he also is said to have said, "Allez en avant, et la foi vous viendra" (Go forward, and faith will come to you).

§4. Vibrating Strings

D'Alembert made significant progress on *the vibrating-string problem:* given the tension in a string and its initial position, figure out how the string moves when it is released. The problem had been studied by Brook Taylor in 1713 [Kli72, p. 478] and by Johann Bernoulli in 1727 [Kli72, p. 479]. Johann's son, Daniel Bernoulli, did related work in 1732 [Kli72, p. 489]. But in 1746 (work published 1747 and 1749) d'Alembert wrote down essentially the modern partial differential equation involved and gave a general solution.

I am discussing the vibrating-string problem for good reason. The vibrations of a string can always be represented as an infinite sum or series of sine waves.[7] Since the initial position of a string can in a sense be arbitrary—we can stretch the string into any shape—it apparently follows that *any* curve, that is, any function, can be represented as a sum or series of sine waves, briefly, a *trigonometric series*. The modern definition of function—of an arbitrary function—evolved as part of the attempt to formulate that conclusion

as a theorem, and set theory evolved more or less as part of the attempt to prove the theorem. The details are an important part of the history I wish to present. In the end it has turned out that the "theorem" is not quite true—there are functions that cannot be represented using trigonometric series—but the ones that can be so represented include functions far stranger than anyone had thought possible.

D'Alembert's general solution to the vibrating-string problem required that the initial curve of the string be periodic, that is, that it repeat the same shape it had on one length of string over and over. He therefore required that the initial position of the string be a periodic analytic expression. That was a substantial restriction on the allowed initial positions of the string: one couldn't start with a string in just any arbitrary configuration, but only one given by a periodic analytic expression.

Shortly after seeing d'Alembert's work, Euler wrote a paper, published in 1748 [Kli72, p. 505], in which he allowed the initial function describing the position of the string to be any function (meeting some other constraints that I omit here) on an interval. He ensured that the function was periodic by simply duplicating its values outside that interval. The function had to be zero at its endpoints since the string was fixed at its endpoints. Thus the duplications matched up at the ends. The function also had to be free of jumps since the string was in one piece. But, and this is the key point, there was no requirement that the function be given by a single (periodic) analytic expression. Euler's broad notion of function from 1734 was now being put to an important mathematical use. In the paper of 1748 he also saw that the motion of the string is periodic in time (the string resumes its initial shape at regular intervals) and that at least some solutions to the problem can be written as sums of sines and cosines.

By 1755 Euler defined a function thus [Kli72, p. 506]: "If some quantities depend on others in such a way as to undergo variation when the latter are varied, then the former are called functions of the latter." He specifically intended to allow functions that are not given by a single equation on the entire domain. In 1763 he wrote to d'Alembert that allowing this more general notion of a function "opens to us a wholly new range of analysis" [Kli72, p. 507].

Daniel Bernoulli, in his work of 1732–1733, was the first to recognize that a string could vibrate at many frequencies—the fundamental frequency that had been studied by Taylor and Daniel's father Johann, and the harmonics (multiples) of that frequency [Kli72, p. 480]. In the early 1740s, Daniel Bernoulli said that a vibrating bar can vibrate at two harmonic frequencies

[7] We add functions or curves by adding them at each point. Thus, the sum $F = f + g$ reduces to an ordinary sum of values at each point: for example, $F(3) = f(3) + g(3)$. Similarly, an infinite sum or series of functions reduces to an ordinary series at each point.

at once. The statement is based on physical understanding, not mathematical derivation. In 1753 he went further: after seeing the work of d'Alembert and Euler, he said that "all the new curves given by d'Alembert and Euler are only combinations of . . . [sinusoidal] vibrations" [Kli72, p. 509]. He was still relying on physics, not mathematics. His claim was that all the new curves can be represented by trigonometric series.

Euler and d'Alembert objected at once to Daniel Bernoulli's claim [Kli72, pp. 509–510]. Euler believed that the functions he had allowed, functions that are defined by different equations in different intervals, could not be a sum of sine functions. A function could not simultaneously be "discontinuous" (given by different expressions in different intervals) and "continuous" (given by a single expression—a sum of sine functions). Moreover, despite his liberal notion of function, a notion that made it possible to piece together a periodic function from nonperiodic ones, Euler argued that since every trigonometric series must be periodic, no nonperiodic function (analytic expression) could be equal to a trigonometric series [GGR72, pp. 245–247]. Euler's attention was now on the analytic expressions themselves. Bernoulli held his ground, and the three continued to disagree with each other through the 1770s [Kli72, p. 513]. Lagrange and the Marquis Pierre Simon de Laplace eventually entered the fray. That all seems a bit bizarre when one sees that Euler (in 1750–1751), d'Alembert (in 1754), and Clairaut (in 1757) had all discovered general methods of representing arbitrary functions by trigonometric series [Kli72, pp. 456–459]. They applied those methods only when they had some (usually physical) reason to believe that a trigonometric-series representation ought to exist. The mathematics did not stand on its own. Indeed, several of the derivations were not correct by our standards. Since nothing could be proved in a reliable way, one tried to confirm the results of a mathematical derivation on some independent grounds. If a result was contrary to expectations, it was often dismissed.

The disagreement about trigonometric series, the paradoxes arising from the use of infinite series, and other disputes created a real internal mathematical need for clarification of the foundations of analysis. The fundamental concepts of function, derivative, and integral had no adequate definition. They had been used in a manner suggested by their applications to simple functions—especially polynomials. As the notion of function was broadened, mainly as a result of work on the vibrating-string problem, that analogical procedure became less and less adequate.

The problem became even more acute with the work of the Baron Jean Baptiste Joseph de Fourier on heat conduction, which involved him in the

problem of trigonometric series. The reception of his work gives some indication of the controversy that surrounded it. In 1807 he submitted a paper to the Academy of Science of Paris. It was rejected by Adrien Marie Legendre, Laplace, and Lagrange. But, to encourage Fourier, the problems he studied were made the subject of an 1812 prize. Fourier's 1811 paper won the prize but was not published. In 1822 Fourier published his great *Théorie analytique de la chaleur.* It included part of the 1811 paper. Fourier became secretary of the Academy two years later, and he had the Academy publish the paper of 1811. (See [Kli72, p. 672].)

Fourier came to see, by a complicated process that we shall not consider, that if for some coefficients b_ν

$$f(x) = \sum_{\nu=1}^{\infty} b_\nu \sin \nu x \quad \text{for } 0 < x < \pi,$$

that is, if the function f could be represented by a trigonometric series[8] in the interval from 0 to π, then for every value of ν

$$b_\nu = \frac{2}{\pi} \int_0^\pi f(s) \sin \nu s \, ds.$$

He took that to mean that the coefficients b_ν of the sum would have to be $2/\pi$ times the area under the curve $f(s) \sin \nu s$ between $s = 0$ and $s = \pi$.[9] As we have seen, formally similar results had been obtained by others, including Euler, d'Alembert, and Clairaut. But Fourier departed from the practice of his day and did not interpret the integral as an inverse differential but geometrically, as an area [GG80a, p. 107]. He observed that the area involved is well defined for an extremely wide variety of fs. Fourier had not been studying vibrating strings, but heat flow, though the mathematics is much the same. His function f represented not the initial position of a vibrating string but the initial distribution of heat in a metal bar. A string must be in one piece, but the temperature in a bar can jump: to produce a bar with a jump in temperature, take a hot bar and a cold bar and join them together. (See [Haw80, p. 152].)

[8] I shall use the term *trigonometric series* to mean a sum of sines, as indicated in the text. Though a general trigonometric series would allow cosines as well, I shall just ignore that fact. The details of what is gained by allowing cosines are irrelevant for our purposes.

[9] Actually, Fourier wasn't quite right. Cantor straightened out the problems in the 1870s, and that work led him to set theory—as we shall see below.

Certainly no single analytic expression need be available for Fourier's temperature function f. The function f might be "freely drawn," and it could even jump about in value The existence of an inverse differential was of no particular concern, and so Fourier was freed from a restriction to analytic expressions. He concluded that *every* function was representable by a special trigonometric series on the interval from 0 to π, the trigonometric series determined by the formula for the b_νs. Such a series is called a *Fourier series*. Fourier tested his opinion by computing the first few b_ν for many fs and plotting the sums of the first few terms of the corresponding sequences. The results looked good.

Summing the trigonometric series above yields a function that is odd and periodic: the values between $-\pi$ and 0 are just the opposites of the values between 0 and π, and the values between $-\pi$ and π are repeated over and over. Thus, the trigonometric series for an *arbitrary* function f, while it may yield the same values as the function between 0 and π, will not yield the same values as the function elsewhere if the function is not odd and periodic. For example, the trigonometric series for the absolute value function (see Figure 3a) has for its sum the "sawtoothed" function in Figure 3b. Fourier stated without fanfare that functions that agree on an interval need not agree elsewhere. That is obviously true on his conception of functions, and it shows how big a change that conception was. For Fourier, in contrast to Euler and Lagrange, functions consist of their values, not the expression used to compute them, and hence may be considered on arbitrarily restricted intervals.

Fourier used his techniques to greatly advance the art of solving partial differential equations. They were too successful to be ignored. Indeed, Siméon-Denis Poisson thought the techniques could be extended to yield general methods to solve all partial differential equations. That did not turn out to be the case, but Poisson did greatly expand their domain of application, and they remain essentially the only available techniques for obtaining exact solutions to partial differential equations subject to boundary conditions [Str87, p. 150]. Since such equations are at the heart of mathematical physics, Fourier's peculiar functions—functions that need not be defined by closed expressions, that could jump about, that did not have the same analytic expression everywhere—became part of the repertoire of every mathematician. In his book, Fourier said [Str87, p. 150], "In general the function $f(x)$ represents a succession of values or ordinates each of which is arbitrary . . . We do not suppose these ordinates to be subject to a common law; they succeed each other in any manner whatsoever." The functions he actually employed were considerably less general. The old basis for analysis had relied on analogies between polynomials and other analytic expressions. It was no longer adequate.

The trigonometric series with coefficients b_ν developed by Fourier—the Fourier series—are infinite discrete series of sine waves. It was natural for the analysts of the day to seek a corresponding integral—to represent a function not as a sum of sines of discrete harmonic frequencies but as a sum of sines of continuously varying frequencies. Results concerning such Fourier integrals were obtained by Fourier (1811), Poisson (1816), and Augustin Louis Cauchy (1816), each of whom was aware of the work of the other two [Kli72, pp. 679–681].

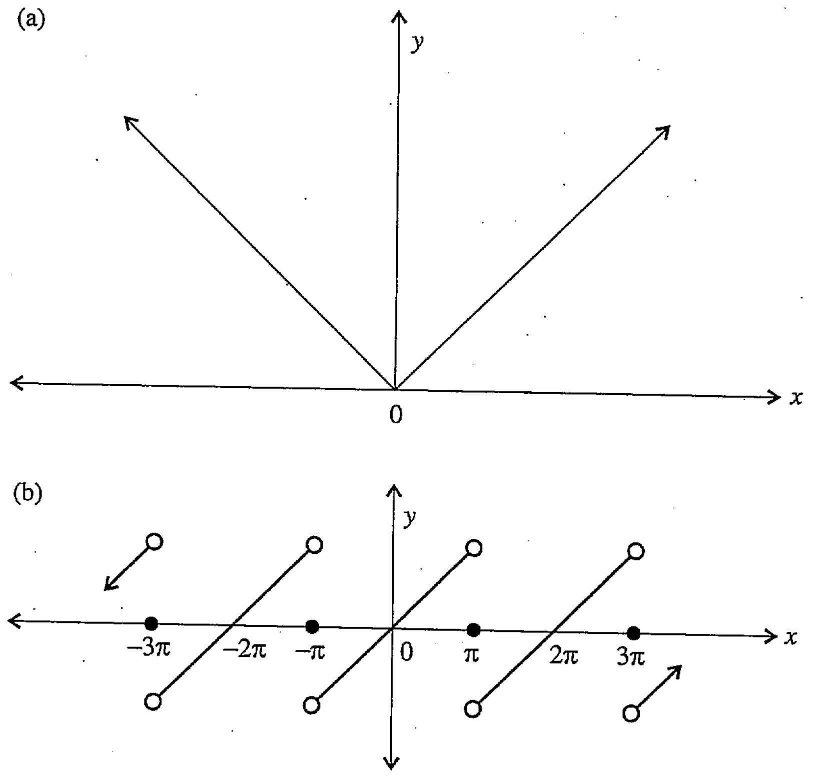

Figure 3. The functions graphed in (a) and (b) agree between $x = 0$ and $x = \pi$. The graph in (b) is obtained by reflecting that piece to obtain the values between $x = 0$ and $x = -\pi$, and then duplicating the graph between $x = -\pi$ and $x = \pi$ for the other values of x. I have given the correct value of the trigonometric series, 0, at the jumps, but Fourier drew vertical lines. (See [GG80a, p. 107] and [Bot86, p. 70].)

§5. Infinity Spurned

It was Cauchy, surely in part stimulated by his work on Fourier integrals, who brought rigor into analysis. His rigor was not rigor for its own sake. Rigor rarely is, as Philip Kitcher has emphasized [Kit83, pp. 213–217].

Cauchy needed a more rigorous development of analysis to continue his research. Euler, who had been interested in series of numbers, could discard the anomalous results divergent series sometimes caused: It was easy enough to test his results by summing a few terms. That procedure is not adequate in the problem context of concern to Cauchy—series of functions, not series of numbers. One would have to sum many terms of a series at many points to show anything, and even then there was the distinct possibility of not trying the right points. Cauchy was prepared to reject some of Euler's techniques to make progress on problems of his own. (See [Kit83, pp. 249–250].) Note that Cauchy's foundational worries primarily concerned series, not infinitesimals. His work was not motivated by the worries of Newton and Berkeley.

The rigor of Bernard Bolzano's *Rein analytischer Beweis des Lehrsatzes, dass zwischen je zwei Werthen, die ein entgegengesetztses Resultat gewähren, wenigstens eine reele Wurzel der Gleichung liege* of 1817 anticipated Cauchy by about four years and in some respects surpassed him. But Bolzano's work was neglected for fifty years. Perhaps that is in part because it came out of a simple desire for rigor, without any subsequent mathematical application [Kit83, p. 264].

Cauchy defined a limit thus:

> When the successive values attributed to a variable approach indefinitely a fixed value so as to end by differing from it by as little as one wishes, this last is called the limit of all the others. (*Cours d'analyse algébrique*, 1821; [Kli72, p. 951].)

He gave essentially the modern definition of what it is for a function to be continuous,[10] not the one current in his day (which was, "defined by a single analytic expression"). He defined the derivative as a limit, and employed only differentials defined in terms of the derivative. He defined convergence and divergence of series and stated baldly that divergent series have no sum. He

developed many useful tests for convergence, and, perhaps most important, he made no use of divergent series. He gave what is now called the Cauchy convergence criterion for a sequence: the sequence $S_0, S_1, \ldots$ has a limit if $|S_{n+r} - S_n|$ is less than any given quantity for every value of r and sufficiently large values of n. He proved the condition necessary but could only assert that it is sufficient on a geometric basis. He found that basis sufficient. A proof of sufficiency had to await the need for one—and a theory of real numbers [Kit83, p. 262].

In his *Résumé des leçons sur le calcul infinitésimal* (1823), Cauchy defined the integral as the limit of a sum of rectangles. That made it possible to make rigorous sense of Fourier's use of integrals of functions that did not have simple analytic expressions, functions for which the integral could not just be defined—in the way that had become usual—as the inverse differential [Haw80, p. 154]. Cauchy gave the first *proof* of the fundamental theorem of the calculus—that the derivative of the integral of a function is the function itself, when the function is continuous. He also discussed the situation in which the function is not continuous. He *defined* length, area, and volume as the values of certain integrals. The old geometric foundation of analysis was turned on its head, and infinitesimals were demoted to a secondary, dispensable role.

Cauchy's definition of the integral is, by modern standards, not very general, and he made false claims. His rigor was apparently adequate for his mathematical purposes, and so he did not pursue rigor further. Many of Cauchy's errors had a single source—Cauchy spoke of the limit of a variable, not the limit of a function, and he did not display the dependence of a variable on the independent variable. He thought of a variable that approached zero as an infinitesimal. His notation was therefore vague with respect to crucial further dependencies. (See [GG80a, p. 121] and [Kit83, pp. 254–255].) Some examples of the false claims: (1) He assumed that a continuous function had a derivative except possibly at a few points. (2) He said that if the sum of a sequence $u_0, u_1, \ldots$ of continuous functions converges everywhere on an interval to a function F then the function F is continuous on that interval. (3) Moreover, he said that under those circumstances one could integrate the function F by summing the integrals of the u_is. The last had been assumed by Fourier in deriving the coefficients b_ν of the Fourier series of a function from the assumption that a trigonometric series for the function exists. For claims (2) and (3) a stronger condition—not convergence but uniform convergence—is required. Nevertheless, Cauchy eliminated the use of diver-

[10] The modern definition of continuity states, roughly, that a continuous function is one with a graph that is free of jumps and wild oscillations, and more precisely that the value of the function at each value of x is the limit of the values of the function at nearby values of x.

gent series, and he launched the effort to give general criteria for the range of applicability of various notions of analysis.

In a letter of 1826, Niels Henrik Abel complained about

> the tremendous obscurity which one unquestionably finds in analysis. It lacks so completely all plan and system that it is peculiar that so many . . . could have studied it. The worst of it is, it had never been treated stringently. There are very few theorems in advanced analysis which have been demonstrated in a logically tenable manner. Everywhere one finds this miserable way of concluding from the special to the general and it is extremely peculiar that such a procedure has led to so few of the so-called paradoxes. [Kli72, p. 947]

In the same year, in a paper on the convergence of binomial series [Kli72, p. 947], Abel praised Cauchy and corrected his claim that the sum of a convergent series of continuous functions is continuous. He gave as a counterexample essentially the sawtoothed function of Figure 3: it is discontinuous, and yet it is the sum of continuous sine curves. He also made some progress toward the required concept of uniform convergence.

Gustav Lejeune Dirichlet, who would be the first to apply the techniques of analysis to obtain results about the natural numbers [Ste88, p. 242], met Fourier in Paris during the years 1822–1825. In 1829 he proved that the Fourier series of any function f meeting a certain condition would always converge to the function. The condition was sufficient but not necessary. It allowed f to have finitely many *exceptional points,* such as bounded discontinuities (for example, jumps). (See [Kli72, p. 966].) Moreover, his proof that the condition was sufficient did not really use the restriction to only finitely many discontinuities. That restriction was only used to ensure that the integrals defining the b_νs were well defined. (See [Haw80, p. 156].)

In the same article Dirichlet concocted the function $f(x)$ that has the value c for rational x and the value d for irrational x [GG80a, p. 126]. It was intended to be an example of a function that could not be integrated and which would therefore not have a well-defined Fourier series—since the coefficients b_ν would not be defined. By 1837, in an article on Fourier series, Dirichlet gave a precursor of the modern definition of a function. That definition is as follows: A function f associates a single value $f(x)$ with each member x of its domain. The association may be perfectly arbitrary—no rule, description, method of computation, or the like is required.[11] Dirichlet said that y is a continuous function of x on an interval if there is a single value of y associated with every value of x on the interval in a continuous manner. Both in the definition of his attempted example of a nonintegrable function and in the definition of a continuous function, Dirichlet clearly viewed a function as being given by its graph alone—by its values—without the need for any associated law. He also extended his sufficient conditions for a function to have a Fourier series to allow more kinds of exceptional points, though still only finitely many. His work posed a problem that is important for our story: can one allow infinitely many exceptional points and still obtain sufficient conditions for the convergence of Fourier series? (See [GG80a, pp. 126–127] and [Bot86, p. 197].)

Dirichlet's standard of rigor exceeded Cauchy's. That is to be expected. As Abel's example showed, Cauchy's techniques were not reliable for deriving general results about the convergence of arbitrary Fourier series. Dirichlet was aware, for example, that the result of taking two limits is sensitive to the order in which they are taken.

Dirichlet's student Georg Friedrich Bernhard Riemann in his *Habilitationsschrift* of 1854 (published posthumously in 1866 by Richard Dedekind) took up Dirichlet's problem. Instead of following Cauchy in showing that certain nice (in today's terminology, piecewise uniformly continuous) functions have well-defined integrals, he gave a definition of an integral more or less like that of Cauchy, but then he sought general conditions, based on an analysis of that definition, under which a function would have an integral. His approach—the chief novelty of which lay in finding general conditions under which functions would be integrable instead of just showing that familiar functions were—was thought for many years to be the most general possible. In particular, he gave an example of a function with infinitely many discontinuities in every interval that could still be integrated in his extended sense. (See [Haw80, pp. 157, 159] and [Haw75].) That work was instrumental in showing that the sort of general definition of an arbitrary function that Dirich-

[11] It is a matter of some dispute how much credit Dirichlet deserves for the modern definition of a function, in part because he restricted his definition to continuous functions. See [You76, pp. 78–79], [Bot86, p. 197], [Vol86, pp. 200–201, 207–209], and [Med91, Chapter 2]. See [You76, pp. 77–80] on the issue of who first actually gave the modern definition of function suggested by Dirichlet's work. I believe Dirichlet defined the notion of a *continuous* function to make it clear that no rule or law is required even in the special case of continuous functions, not just in general. That would have deserved special emphasis because of Euler's *definition* of a continuous function as one given by single expression—or law. But I also doubt that there is sufficient evidence to settle the dispute.

let's ideas suggested has a genuine mathematical point: one could prove interesting things about classes of arbitrary functions that included "pathological" ones. Riemann raised many problems concerning Fourier series, including the one studied by Cantor that led to set theory. (See below.) Some of the problems are still unsolved and of great interest. (See [GG80a, pp. 132, 138].)

It was Karl Weierstrass who developed the methods needed to attack Riemann's problems. Most of analysis in the latter part of the nineteenth century consisted of applying Weierstrass's methods to Riemann's problems. (See [GG80a, p. 132].)

Weierstrass did important research on the representation of functions via power series. In the process, he put analysis into its modern rigorous form during the years 1841–1856, when he was a high-school teacher. That work did not become known until the late 1850s, when he finally obtained a professorship. Thus, for example, though Weierstrass understood the uniform convergence of series by 1842 (that was necessary for his work on power series), the concept actually became known through the work of George Gabriel Stokes and Ludwig Philipp Seidel, who in 1847 and 1848 independently arrived at formulations of closely related notions that are not quite as generally useful. Seidel was a student of Dirichlet. Dirichlet was aware that a convergent sum of continuous functions (sines) could be discontinuous, contrary to Cauchy's stated result. That is what got Seidel started. (See [GG80a, pp. 127–128] and [Med91, pp. 88–91].) Cauchy corrected himself in 1853. But he did not pursue the matter even to the extent of seeing where else he had illicitly assumed uniform convergence earlier. (See [Bot86, pp. 207–208].)

Weierstrass was one of the many mathematicians who corrected Cauchy's belief that continuous functions have derivatives except possibly at a few points by giving examples of functions that are continuous but fail to have derivatives at many points. Weierstrass's example was continuous everywhere but differentiable nowhere.

Weierstrass replaced Cauchy's definition of a limit that involved such notions as "approach indefinitely" and "differ by as little as one wishes" by the following definition, in which "as little as one wishes" has become ϵ: a function $f(x)$ has limit L at $x = x_0$ if for every $\epsilon > 0$ there is a δ such that if $|x - x_0| < \delta$ and $x \neq x_0$, then $|f(x) - L| < \epsilon$. The "variable," that is the function, in Cauchy's terminology "ends" in the interval defined by $|x - x_0| < \delta$ by differing from the "fixed value" L by "as little as one wishes," that is, by less than ϵ. The replacement enabled Weierstrass to distinguish convergence from uniform convergence and to make allied distinctions, cleanly and naturally. That was at least part of the motivation for the new rigor [Kit83, p. 257].

Weierstrass also gave a theory of irrational numbers around 1860 [Kli72, p. 979]. Dedekind had developed a theory in 1858 [Ded72, p. 2]. Their predecessors for the most part, when they defined irrational numbers at all, had defined them as certain limits of sequences of rational numbers. That procedure, as Weierstrass's precise definition of a limit makes clear, does not suffice: As Cantor pointed out in 1883, the number L must already exist for it to be the limit of a sequence. If we start with only rational numbers, a sequence that "converges to an irrational number" will not have a limit L.

In the early 1830s, William Rowan Hamilton gave a different treatment of the irrationals, taking time as a basis [Kli72, p. 983]. Weierstrass would not have approved. Weierstrass defined a variable simply as a letter that may be assigned various values. He banished the old idea of a variable quantity, which often in some metaphorical sense varied with time.

As Hilbert put it in 1925,

> As a result of his penetrating critique, Weierstrass has provided a solid foundation for mathematical analysis. By elucidating many notions, . . . he removed the defects which were still found in the infinitesimal calculus . . . If in analysis today there is complete agreement and certitude in employing the deductive methods which are based on the concepts of irrational number and limit, and if in even the most complex questions of the theory of differential and integral equations . . . there . . . is unanimity with respect to the results obtained, then this happy state of affairs is due primarily to Weierstrass's scientific work. [Hil26, p. 183]

Cauchy and Weierstrass had eliminated time, infinitesimals, and infinite quantities from the foundations of analysis, and by so doing they made possible a standard of rigor surpassing that of the Greeks.

§6. Infinity Embraced

In 1817 Bolzano tried to prove that a continuous function that is both negative and positive in an interval takes on the value zero in that interval. He made use of the fact, which he also tried to prove, that every bounded set of values has a least upper bound. An adequate proof had to await an adequate theory of the real numbers. Weierstrass used his own theory of the irrationals and techniques suggested by those of Bolzano to prove in the 1860s that every bounded infinite set of points has a *limit point*—that is, a point such that every interval around it contains infinitely many members of the set. (For example, 1 is a limit point of the set $\{0, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \ldots\}$. Intuitively one sees that the members of the set crowd against 1.) The result is now called the Bolzano–Weierstrass theorem. (See [Kli72, p. 953].)

In the years 1869–1872, Charles Méray, Cantor, Heinrich Eduard Heine, and Dedekind all published theories of the irrational numbers [Kli72, p. 983]. Cantor's was a modification of Weierstrass's earlier theory [Jou15, p. 26]: Weierstrass defined real numbers in terms of series of rational numbers, while Cantor used sequences. Dedekind published his theory in response to Cantor's publication [Dau79, p. 48]. In 1886 Stolz showed that one can identify the irrationals with the nonrepeating decimals [Kli72, p. 987]. Every single one of the theories of irrationals defines irrationals in terms of some actually infinite sets or sequences. A nonrepeating decimal involves an infinite sequence of digits. Dedekind's theory of cuts defines $\sqrt{2}$, for example, in terms of the infinite set of all positive rational numbers p such that $p^2 > 2$. (That set and the set of remaining rational numbers—that is, those rational numbers p that are negative or such that $p^2 < 2$—cut the rational numbers into two parts, an initial segment and a final segment, hence the name "cut.") Cantor's theory of Cauchy sequences defines a real number to be associated with an infinite set of infinite sequences of rational numbers. And so forth.

Dedekind's theory closely resembles that of Eudoxus for incommensurable ratios. Roughly speaking, the upper and lower parts of the cut correspond to the commensurable ratios greater than and less than a given incommensurable ratio. Indeed, Dedekind gave credit to Book V of Euclid's *Elements*. Cantor felt his own theory to be superior to that of Dedekind because it makes use of sequences of rational numbers—objects familiar to analysts—instead of the unfamiliar "cuts" [Kli72, p. 986].

The definitions of the irrational numbers provide one of the great ironies in the history of mathematics: Cauchy and Weierstrass had eliminated infinitely small and infinitely great numbers from analysis and replaced them by limits. But the theory of limits that thereby became so central required a clearer theory of the real line, that is to say, a theory of the irrational numbers. And that theory promptly reintroduced the infinite into analysis. The old infinity of infinitesimal and infinite numbers was simply replaced by the new infinity of infinitely large collections.[12]

[12] See Russell's *Principles of Mathematics* [Rus03, p. 304] for a related sentiment.

In 1831 Carl Friedrich Gauss said [Kli72, p. 994], "I protest against the use of an infinite quantity as an actual entity; this is never allowed in mathematics. The infinite is only a manner of speaking, in which one properly speaks of limits to which certain ratios can come as near as desired, while others are permitted to increase without bound." But only 52 years later, we find this in Cantor's *Grundlagen* [Can76, p. 75]: "The idea of considering the infinitely large not only in the form of the unlimitedly increasing magnitude and in the closely related form of convergent infinite series . . . but to also fix it mathematically by numbers in the definite form of the completed infinite was logically forced upon me, almost against my will since it was contrary to traditions which I had come to cherish in the course of many years of scientific effort and investigations."

Fourier, among others, had given a proof that if a function is representable by a trigonometric series, then the series is unique, that is, that any two trigonometric series that converge to the function are the same. We exhibited the main parts of such a proof above: If there is a series, its coefficients must be the b_νs we gave, and the series is therefore unique. The proof does not work, because the formula for the b_νs was obtained by integrating the series term by term. As noted earlier, even Cauchy believed such a procedure legitimate, but it works only if the series is uniformly convergent.

Weierstrass emphasized the importance of uniform convergence, and Heine became interested. He may have learned about it from Cantor, who had studied with Weierstrass before becoming Heine's colleague at Halle. In a paper of 1870, Heine noted the gap in the proof of the uniqueness of a trigonometric expansion. What had actually been proved was that if a function has a uniformly convergent trigonometric series, then that series is the Fourier series and it is the only uniformly convergent trigonometric series that sums to the function. But it was known that even Fourier series need not be uniformly convergent. The Fourier series for the sawtoothed function discussed earlier provides an example. Heine gave some positive results concerning uniqueness.

Influenced by Heine, Cantor proved that the trigonometric-series representation of a function is unique. Uniform convergence is not required. That result applies to a trigonometric series that converges everywhere.

Cantor began to extend his result to allow exceptional points. In 1871 he showed, verifying a belief of Riemann, that if two trigonometric series converge to the same function everywhere except possibly at finitely many points, then that is enough to ensure that they are the same series.

In 1872 Cantor obtained results that allowed infinitely many exceptional points, answering a question of Riemann: He defined the derived set S' of a set S of real numbers to be the set of limit points of S. For example, if S is the set $\{0, \frac{1}{2}, \frac{3}{4}, \frac{7}{8}, \ldots\}$, then S' is the set $\{1\}$ whose only member is 1. One can form the derived set of a derived set, that is, a second derived set, and so forth. Cantor proved a generalization of his earlier result: Suppose a set S of real numbers is such that for some n the nth derived set is finite. If two trigonometric series converge to the same function except possibly at points in S, then they are the same. Cantor gave his definition of the real numbers in the paper in which the generalization appeared. He needed it to show that for every n there is a set S whose nth derived set is finite and nonempty while the $n-1$st derived set is infinite. That is, he needed it to show that allowing the iteration of his derived-set operation actually led to new possible sets S of exceptional points.

Cantor's paper was one of the first in which infinite sets of points received careful, explicit consideration. Dirichlet, in 1829, had proposed a condition for a function to be integrable that was a condition on the set of points of discontinuity of the function, a condition that was clearly concerned with infinite sets of points of discontinuity. The condition is worth stating, since it will come up again later. It is that the set of points of discontinuity be *nowhere dense*—that is, that within every interval there is contained an interval that includes no points of discontinuity. The intuition is supposed to be that nowhere-dense sets are in some sense small. But Dirichlet published no results [Haw80, p. 156]. In his doctoral thesis of 1864, Rudolph Lipschitz developed a condition under which a function would have a convergent Fourier series even when it had infinitely many points of oscillation. But the proof made substantial use of the structure of trigonometric series [GG80a, p. 137]. In 1870 Hermann Hankel, who had studied under Riemann, developed a condition under which a function would be integrable (in Riemann's sense) that involved sets of points of discontinuity. But his primary focus was on integrability [Haw80, p. 166].

Cantor's theorem of 1872, unlike the other results just mentioned that had involved infinite sets of points, was proved using nothing new other than a careful study of the structure of the relevant sets of points. The only other ingredient was an absolutely straightforward application of his earlier results concerning trigonometric series. Moreover, though Cantor used only finite iterations of the derived-set operation in 1872, he was already aware of the possibility of infinite iterations: Given a set P, let P' be its derived set, and in

general let $P^{(k+1)}$ be the derived set of $P^{(k)}$. We so far have the sequence

$$P^{(0)} = P,\ P^{(1)} = P',\ P^{(2)} = P^{(1)\prime} = P'',\ P^{(3)}, \ldots.$$

Now let $P^{(\infty)}$ be the set of points that are in $P^{(k)}$ for every finite k—the points that the operation of taking a derived set has not yet eliminated. Then we can continue

$$P^{(0)},\ P^{(1)}, \ldots,\ P^{(\infty)},\ P^{(\infty+1)} = P^{(\infty)\prime},\ P^{(\infty+2)}, \ldots,\ P^{(\infty\cdot 2)}, \ldots,$$
$$P^{(\infty\cdot 3)}, \ldots,\ P^{(\infty^2)}, \ldots,\ P^{(\infty^3)}, \ldots,\ P^{(\infty^\infty)}, \ldots,\ P^{(\infty^{\infty^\infty})}, \ldots.$$

(See [Dau71, pp. 211–213].) Cantor discovered the sequence in attempting to analyze the structure of complicated sets of real numbers, not, as is sometimes said, by attempting to generalize the natural numbers. That is not to say that he did not come to view the sequence as a generalization of the natural numbers. He did. But the sequence arose from attempts to consider complicated functions on the real numbers (those with complicated sets of exceptional points), not from a study of the infinite or an attempt to generalize the natural numbers. The modern theory of the infinite evolved in a contiguous way out of the mathematics that preceded it. Though I disagree with Kitcher in detail about the origins of Cantorian set theory [Kit83, p. 207], I fully endorse his main thesis that new mathematics by and large—and in particular Cantorian set theory—evolves out of old mathematics [Kit83, Kit88].

Cantor had obtained mathematical results by focusing on the structure of infinite sets of points, and he knew that there were sets of points with more complicated structures. (His theorem, recall, concerned only sets P such that $P^{(k)}$ was finite for some finite k—no use was made of sets such that, say, $P^{(\infty^2+6)}$ was infinite, but $P^{(\infty^2+7)}$ was finite.) Understanding more complicated sets was connected with understanding more complicated and arbitrary functions—ones with more complicated sets of exceptional points. Cantor made the fateful decision to turn his attention to the study of sets of points in their own right.

III

Sets of Points

§1. Infinite Sizes

Cantor began by studying the two most obviously interesting sets of points: the set of rational numbers and the set of real numbers. He looked for differences between the two that are relevant to the fact that the real numbers are continuous while the rational numbers are not. In 1874 he published a paper in which he demonstrated the remarkable result that the algebraic numbers (and hence their subset, the rational numbers) can be placed into one-to-one correspondence with the natural numbers, while the real numbers cannot. The set of rational numbers is thus shown to have the same size as the set of natural numbers—they can be paired off—but the set of real numbers is shown to be bigger than the set of rational numbers. The proof that Cantor gave that the real numbers cannot be placed into one-to-one correspondence with the natural numbers is not the one that is most familiar today. In particular, it did not show that there were any other infinite sets that were of other sizes, and Cantor knew of nothing larger than the set of real numbers. It follows from Cantor's results that one cannot define the real numbers in terms of finite sets of rational numbers: there aren't enough finite sets of rational numbers. The use of actual infinity in the various definitions of the irrational numbers had been no accident.

Cantor began to investigate whether he could put the points on a plane into one-to-one correspondence with the points on a line, perhaps as part of a search for larger infinite sizes. In 1878 he published the unexpected result that one can indeed put the points on a plane, that is, a two-dimensional space, or indeed any n-dimensional space, into one-to-one correspondence with the points on a line. The techniques he used suffice to show that the points of an ∞-dimensional space can be put into one-to-one correspondence with the points on a line. Cantor also stated that he could show that every infinite set of points on the line could be placed into one-to-one correspondence with either the natural numbers or the real numbers—that there are no intermediate possibilities [Jóu15, p. 45]. His proof turned out to be incorrect, and a proof of his claim, known today as the continuum hypothesis,[1] continued to elude Cantor, though much of his work was apparently motivated by attempts to prove it. The problem is still open today. It has been shown that the truth or falsehood of the continuum hypothesis cannot be settled on the basis of the set-theoretic principles we accept today (assuming they are consistent): In 1938, Kurt Gödel showed that the continuum hypothesis cannot be disproved on that basis [Göd90, p. 26]. In 1963, Paul J. Cohen showed it cannot be proved on that basis (see [Jec78, p. 176]). It therefore could not have been settled on the basis of the similar principles that Cantor employed.

Since spaces of different dimensions can be placed into one-to-one correspondence, Cantor's work posed the problem of seeing how spaces of different dimensions differ. Dedekind observed that spaces of different dimensions cannot be placed into one-to-one correspondence by a *continuous* function—Cantor's correspondences were discontinuous. Luitzen Egbertus Jan Brouwer was the first to provide a satisfactory proof of Dedekind's observation, in 1911 [Dau80, p. 188].

Leopold Kronecker, Cantor's former teacher, was an editor of the journal to which Cantor submitted his paper on dimension. Kronecker believed that all of mathematics should be based on the natural numbers. That is an early version of a view that we shall refer to as *finitism*. He also believed that every definition of a property should give a method for determining whether or not an object has the property. That is an early version of a view that we shall refer to as *constructivism*. Note that, although Kronecker was both a finitist and a constructivist, there is no logically necessary association between the two: one can be a finitist without being a constructivist, or vice versa. That point requires emphasis since finitism and constructivism are so frequently associated that they are often not clearly distinguished. When Cantor's paper did not appear immediately, Cantor suspected Kronecker of delaying it ([Dau80, pp. 188–189], but see also [Edw88]).

In 1879 Cantor published the first of a series of papers about subsets of the real line. In that series and in related papers he defined many notions still

[1] The source of the name is Felix Bernstein's Ph.D. dissertation, which discussed the "continuum problem." See [Moo82, p. 56].

in use today concerning subsets of the real line and other spaces. Though I have modernized the notation, the following definitions are Cantor's: A set P is *everywhere dense in the interval* (a, b)—that is, in the set of real numbers greater than a and less than b—if $(a, b) \subseteq P'$, where P' is the set of limit points of P. A set P is *perfect* if $P = P'$. A set P is *isolated* if $P \cap P' = \varnothing$. A set P is *closed* if $P \cap P' = P'$. I give the definitions to emphasize that the papers are very much concerned with the real line, since I am—somewhat misleadingly so far as historical summary is concerned—going to concentrate on the aspects of them that led to set theory in a more abstract form. Cantor's research centered around two ideas—that of the derived set and that of the transfinite symbols. I am going to concentrate on the work connected to the transfinite symbols. For more details of the other aspect of Cantor's work, and references, see [Dau79].

§2. Infinite Orders

In 1879 Cantor defined two sets to be of the *same power* if they can be placed into one-to-one correspondence. He noted that the concept generalizes that of whole number, and that power "can be regarded as an attribute of any *well-defined* collection, whatever may be the character of its elements." In 1880 Cantor published his transfinite symbols for iteration of derived sets—$\infty^{\infty^3} + 1$, and so forth—for the first time. He said, "we see here a dialectic generation of concepts which always continues further and thus is free of any arbitrariness."

By 1882 the "symbols" were an object of study in their own right in *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* (Foundations of a general theory of manifolds, published in 1883 [Can83]). To separate his transfinite ordinal numbers from the notion of increasing without bound symbolized by ∞ in analysis, he began to use ω instead of ∞. The use of the symbol[2] ω has been standard ever since. Cantor introduced what was to become the distinction between cardinal and ordinal numbers: The sets $(a_1, a_2, \ldots)$ and $(b_2, b_3, \ldots, b_1)$ have the same power, or cardinality, but their numberings, their orders, are different.[3] The first one has the order ω, while the second has the order $\omega + 1$. Indeed, the very same set can be numbered or counted in more than one way: consider $(a_1, a_2, \ldots)$ and $(a_2, a_3, \ldots, a_1)$. If a set is finite, there is only one order to give it even though one can vary which element of the set occurs at which point in the order, and so finite ordinal and cardinal numbers coincide. Cantor defined operations of addition and multiplication on the ordinal numbers, which is part of what justified thinking of them as *numbers* [Kit83, p. 174].

[2] The symbol ω, a lower-case omega, is the last letter of the Greek alphabet.

[3] The notation used to indicate orders does not meet modern standards of rigor. It is nonetheless clear enough what is meant. I use it here and below in this chapter since it is indicative of the way in which Cantor thought about orders.

In the *Grundlagen* Cantor declared for the first time that there are many infinite sizes: He showed how to produce a set of power greater than the natural numbers, namely, the set of all ordinal numbers of the power of the natural numbers. (As illustrated above, ω and $\omega + 1$ are such ordinal numbers.) The proof he offered is a straightforward generalization of the one he used to show that there are more real numbers than rational numbers. (The two proofs are given and then compared in detail in §IV.2.) He called the power of the natural numbers (I), the new power, (II). The power (III) is the power of the set of all ordinal numbers of power (II). And so forth. He said that for every ordinal number γ there is a new power (γ). He did not have full control of all of the details, but with hindsight we can see that the proof was essentially correct. The construction starts with the natural numbers. It is an iteration of much the same sort that led to the "symbols" for successive derived sets. Indeed, as Philip E. B. Jourdain said, it may have been the primary reason for Cantor to have considered the "symbols" on their own, separately from derived sets [Jou15, p. 51].

The proof that the powers are distinct provides no way to make contact with the power of the real numbers. For all Cantor knew, the powers he had constructed were all smaller than that of the real numbers, or even all incomparable with that of the real numbers. Cantor made an additional assumption in the *Grundlagen* that guaranteed that the new powers were comparable with that of the real numbers: he assumed that the real numbers form a set and that they can therefore be well ordered.[4] That ensures that the power of the real numbers is less than, equal to, or greater than each of the new powers, but it gives no information about which. Cantor felt less than certain about the new

[4] As we shall discuss in detail below, in §IV.2, Cantor had good reason for thinking he would eventually be able to prove that. The formulation in the text here is in one respect a bit misleading: as we also discuss in §IV.2, for Cantor being well-orderable was in effect constitutive of being a set, and so the right way to put the point would be, "He assumed that the real numbers can be well ordered and that they therefore form a set."

assumption. Indeed, at one point Cantor worried briefly that the power of the continuum was not comparable with any of the infinite powers (γ) [Hal84, pp. 42, 73, 76–77].

During the period in which Cantor was working out the theory of ordinal numbers, many mathematicians, including Cantor himself, came up with examples of nowhere-dense (§II.6) subsets of an interval of real numbers that are not small in an important sense: They cannot be covered by finite unions of intervals of arbitrarily small total length. That is, for example, there is a nowhere-dense set P such that for any n if P is contained in a union $[a_1, b_1] \cup \cdots \cup [a_n, b_n]$ of n intervals, then the sum $|a_1 - b_1| + \cdots + |a_n - b_n|$ of the lengths of the intervals is greater than 1. It seems reasonable in some sense to say that P has length at least 1, and that it is therefore not small despite the fact that it is nowhere dense. The appropriate sense is called *outer content*. Outer content was introduced by Cantor and, independently, by Stolz in 1884 [Haw80, p. 168]. As we shall see in §3, outer content was to be important in devising a notion of integral sufficiently general for the purpose of studying Fourier series. Thus, set theory was not only the product of problems within analysis; it also gave rise early on to fruitful new ideas for solving problems within analysis.

In 1885 Cantor prepared a paper in which he studied general linear orders, defined independently of the rational numbers or the real numbers.[5] The paper was not published in Cantor's lifetime, but it is worth mentioning from our perspective, since it was probably the earliest study of abstract structure independent of some familiar intended mathematical model. Another likely candidate is Dedekind's work that led to [Ded88]. Since Cantor and Dedekind were frequent correspondents, it would be worthwhile to know more about their discussions related to the present topic.

In the 1885 paper Cantor said that pure mathematics is just set theory, in the sense that all of mathematics can be understood in purely set-theoretic terms [Dau80, p. 202]. The year before Gottlob Frege had published *Die Grundlagen der Arithmetik*, deriving arithmetic from logical principles. The later development of Fregean foundations of arithmetic, along with Dedekind's work, was part of what led to the eventual widespread acceptance of Cantor's insight.

I have been describing in detail the extent to which set theory is intertwined with analysis, particularly with the theory of trigonometric series, but there is another sense in which set theory has become important to mathematics: Mathematics is today thought of as the study of abstract structure, not the study of quantity. That point of view arose directly out of the development of the set-theoretic notion of abstract structure.

The motives of Dedekind and Frege were rather different from those of Cantor: Dedekind was seeking to give a foundation of arithmetic "entirely independent of the notions or intuitions of space and time" [Ded88, p. 31]. Frege was studying logic as part of a philosophically motivated program of giving an explicit foundation for arithmetic—the development of logic was needed to ensure that no assumptions were going unnoticed. Both wished to provide a foundation for analysis. It is arguable that each is an exception to the usual rule that rigorization is not undertaken for its own sake. Even if they are exceptions, they are not harmful ones: they had ample motivation for increased rigor in the recent great success of the rigor of Cauchy and Weierstrass, and in their perception that the program of arithmetization of analysis advocated by Dirichlet [Ded88, p. 35] and Weierstrass had not been completed.

It is important to realize that though rigor and the systematization of analysis were the motives of Dedekind and Frege, they were not Cantor's motives, despite the usual account. Cantor was studying sets of real numbers for mathematical reasons that grew out of the study of the Fourier series of increasingly arbitrary functions. He did not work axiomatically. He believed in the reality of his ordinal numbers and sets, and he saw himself as discovering their properties. Therefore, no axioms were necessary. The fact that Cantor did not work axiomatically shows that, in contrast to Dedekind and Frege, he did not see his project as that of working out the consequences of a system of assumptions or as that of systematizing a body of knowledge. When a fact seemed obvious or elementary, Cantor just stated it without proof.

In Cantor's *Grundlagen*, powers were not associated with cardinal numbers. In 1883 Cantor made such an association for the first time. In 1886 he introduced a notation for cardinal numbers, and in 1887 he gave definitions of operations of addition and multiplication for cardinal numbers. The ordinal numbers came first, and they were always more important than the cardinal numbers for Cantor. (See [Dau79, pp. 179–181].)

[5] A *linear order* is a set M with a binary relation $<$ on it such that no m in M is such that $m < m$ (irreflexive); for all l, m, and n in M, if $l < m$ and $m < n$, then $l < n$ (transitive); and for all m and n in M, $m < n$, or $m = n$, or $n < m$ (connected). Examples include the natural numbers, the rational numbers, the real numbers, the ordinal numbers, or any subset of any of those, in each case with the obvious order relation.

In 1888, Dedekind published his theory of the natural numbers [Ded88]. A year later, Peano, with reference to Dedekind's work, gave a semiformal version of what has become the standard axiomatization of the natural numbers.[6] Peano's paper gave the first statement of the need to distinguish between the relation of set membership (for which Peano introduced the symbol ∈) and set inclusion [Pea89, p. 86]. Peano also introduced what has become known as a comprehension principle ([Pea89, p. 90], see also [Ken80, p. 26]): every "condition" (that is, every "proposition containing the indeterminate x") determines a class, namely the "class composed of the individuals that satisfy [the] condition." Peano also gave foundations for the rational and irrational numbers, and even discussed Cantor's set theory. In 1890, Peano defined a continuous curve that hits every point in the unit square at least once. He also introduced the distinction between an individual and the class composed of that individual alone, and denied that one can select members from infinitely many classes without a determinate rule. (See [Kli72, pp. 988, 1018], [Ken80, p. 33], and [Moo82, p. 76].)

In 1891, Cantor published his diagonal argument.[7] It yielded a new proof that there are more real numbers than natural numbers, and, much more important, it was the first completely worked out argument that showed that there are infinitely many infinite powers. Indeed, it shows more: it shows that given any set, there is another of greater power. By applying that fact to the set of real numbers, Cantor showed for the first time that there is an infinite power strictly greater than that of the set of real numbers.

In 1892 Frege published a review of Cantor's 1887 paper, the paper that had introduced cardinal arithmetic. He fully endorsed Cantor's acceptance of the actual infinite and saw that Cantor's work had important consequences for analysis. Nonetheless, he devoted the bulk of the review to castigating Cantor

[6] Peano arithmetic, PA, will be a useful example in later chapters. I shall therefore describe a convenient version. It is not exactly the one given by Peano. Here it is: Every number has a successor. The number 0 is no number's successor. Numbers with the same successor are equal. The sum of any number x and 0 is x. For any number x, the sum of its successor and any number y is the successor of the sum of x and y.

The product of any number x and 1 (the successor of 0) is x. For any number x, the product of its successor and any number y is the sum of the product of x and y with y. Any property that holds of 0 and that is such that if it holds for any number x then it holds for the successor of x, holds for every number. The last axiom listed is known as the induction axiom.

[7] The paper is translated as Appendix B to Chapter IV.

for his reliance on intuition and an ill-defined notion of "abstraction." He advocated logical rigor and explicit definition. He had, however, no doubt that Cantor's theory could be developed in a satisfactory manner. One year later, the first volume of Frege's *Grundgesetze der Arithmetik, begriffsschriftlich abgeleitet* was published. In it he began to carry out in detail the program of his *Grundlagen*: developing arithmetic within a formal system, giving fully formal proofs to show that nothing was being smuggled in on the basis of intuition. In the *Grundgesetze,* Frege introduced a theory that brought his program into contact with Cantor's theory of sets. Presumably that theory was what Frege thought was the appropriate background in which to develop Cantorian set theory in a rigorous manner. We shall be discussing that and related issues in some detail in the next chapter. The second volume of the *Grundgesetze* (published in 1903) contains a theory of cardinal numbers. (See [Dau79, pp. 220–225] and [Dum67].)

§3. Integration

In 1892 Camille Jordan gave a definitive formulation of the Cauchy-Riemann integral. The integral of a function in two dimensions—a surface integral—was at the time usually defined over the region bounded by a closed curve. The surface integral was defined in terms of limiting values of sums taken over arbitrary partitions of the plane into rectangles. There was an obvious problem about what to do with rectangles on the boundary of the region—should rectangles neither wholly within nor wholly outside of the region of integration be included or excluded from the sums? That problem was finessed by claiming that the sum of the areas of the rectangles on the boundary went to zero in the limit and that it therefore didn't matter whether the rectangles on the boundary were included or excluded. But Peano's curve—which had every point in a region on the boundary—suggested that the claim was in trouble.[8]

Jordan solved the problem by moving from rectangles to subsets of the plane, in a Cantorian sense. The Cantor–Stolz notion of outer content extends

[8] See [Haw75] for a detailed chronicle of the modern theory of integration, including Cantor's contribution and the strong impact of set theory on the development of the theory. The history of integration presented here is abbreviated and simplified in important respects, since my only aim is to give an example of how Cantor's set theory influenced the subsequent development of analysis.

in an obvious way from the line to the plane: just use rectangles instead of intervals. Jordan introduced a notion of inner content suggested by that of outer content. The outer content of a set is defined using the areas of finite sets of rectangles that contain the set. It is the greatest lower bound of such areas. The inner content of a set is the least upper bound of the areas of finite sets of pairwise disjoint rectangles that are contained within the set. Jordan said that a subset of the plane is *measurable* if it has outer content equal to its inner content. Naturally the *measure* of a measurable set is its outer, or, indifferently, its inner content. One easily sees that the measures of familiar sets of points on the plane are just their areas.

Now any measurable set has a well-defined "area" or measure, and so one no longer needs to have rectangles play a special role. Jordan defined the integral in terms of limiting values of sums over arbitrary partitions of the plane into measurable sets instead of over partitions into rectangles. It was then natural to allow the region of integration to be an arbitrary measurable set, instead of just the interior of a curve. Jordan showed that a set is measurable if and only if the outer content of its boundary is zero. That is what he needed to show that the sum of the areas on the boundary went to zero in the limit, and that the integral was therefore well defined. In 1893 Jordan incorporated his approach into his text, the *Cours d'analyse*. The next generation of French mathematicians learned Jordan's set-theoretic formulation of analysis. (See [Haw80, pp. 169–171].)

In 1895 Cantor defined cardinal exponentiation and observed that the power of the set of real numbers is $2^{\aleph_0}$. He could then have given the algebraic formulation of the continuum hypothesis that is standard today: $2^{\aleph_0} = \aleph_1$. In fact he did not. (The symbol $\aleph_0$, aleph naught, denotes the cardinal number of the set of natural numbers—(I) in the old notation. The symbol $\aleph_1$ denotes the next larger cardinal number, (II). And so forth. In particular, $\aleph_\omega$ denotes the ωth cardinal number.)

In 1902 Henri Lebesgue, building on important intermediate work of Emile Borel and others, built set theory into the very foundations of analysis. He changed the definition of outer content of subsets of the unit interval to allow actually infinite *denumerable* sets of intervals instead of just finite ones. (I have switched from the plane to the unit interval for simplicity—his definitions on other domains are based on the one sketched here.) That is, he allowed $[a_1, b_1], \ldots, [a_n, b_n], \ldots$ in addition to $[a_1, b_1], \ldots, [a_n, b_n]$.[9] He then defined the inner content of a subset E of the unit interval to be 1 minus the outer content of the complement of E. The corresponding notion of measurability is known as Lebesgue measurability.

The Lebesgue integral may be defined in just the same way that Jordan defined the Cauchy-Riemann integral, except using Lebesgue measure instead of Jordan measure. The Lebesgue integral has many convenient properties that the Cauchy-Riemann integral does not. For example, if a sequence of Lebesgue integrable functions on a set of finite measure is uniformly bounded (which just means that there is a B such that all of the values of all of the functions are less than B) and converges to a function, then that function is Lebesgue integrable and the value of the integral is the limit of the sequence of values of the integrals of the functions:

$$\int_a^b \lim_{n\to\infty} f_n(x)\,dx = \lim_{n\to\infty} \int_a^b f_n(x)\,dx.$$

The Lebesgue integral has many applications in the theory of Fourier series. For example, Lebesgue showed in 1903 that if a bounded function is represented by a trigonometric series, then the series must be the Fourier series (where, of course, the Fourier coefficients are now defined using Lebesgue integration). (See [Kli72, pp. 1044–1048] and [Haw80, pp. 172–179].)

In 1915 Jourdain translated Cantor's *Beiträge zur Begründung der transfiniten Mengenlehre* into English (published as *Contributions to the Founding of the Theory of Transfinite Numbers*). He replaced *sets* [*Mengen*] in the title by *numbers*, "since these memoirs are chiefly occupied with the investigation of the various transfinite cardinal and ordinal numbers and not with investigations belonging to what is usually described as . . . 'the theory of sets' . . . —the elements of sets being real or complex numbers which are imaged as geometrical 'points' in space of one or more dimensions" [Can15, p. v, preface].

§4. Absolute vs. Transfinite

Cantor's study of sets began with his work on arbitrary functions and the discovery of the transfinite symbols, and it always remained tied to that beginning. Cantor believed he had discovered that between the finite and the "Absolute," which is "incomprehensible to the human understanding," there is a third category, which he called the *transfinite*. Cantor's initial reasons for postulating the Absolute were primarily theological, and theology continued

[9] As exemplified in the text, a *denumerable* set is one that can be indexed by the natural numbers.

to play an important part in his notion of the Absolute throughout his life. We shall discuss some of his mathematical ideas about the Absolute below, but our primary focus will be on Cantor's understanding of the transfinite. The Absolute enters into the discussion largely so that we can see how Cantor contrasted it with the transfinite. In the *Grundlagen* of 1883 he said that "the Absolute can only be acknowledged and admitted, never known, not even approximately," and that he was convinced "that the domain of definable quantities is not closed off with the finite quantities, and that the limits of our knowledge may be extended accordingly without this necessarily doing violence to our nature." In 1887 he characterized the transfinite as "in itself constant, and larger than any finite, but nevertheless unrestricted, increasable, and in this respect thus bounded." Cantor, from the beginning, devoted his efforts to understanding only the increasable infinite. (See [Hal84, pp. 13, 14].)[10]

In fact, Cantor's notion of the transfinite is even more specific, as he makes clear in the *Grundlagen:*

> The assumption that, besides the Absolute, which is unreachable by any determination, and the finite, there are no modifications which I call actually-infinite, that is to say which are determinable through numbers—this assumption I find to be quite unjustified . . . What I assert and believe to have demonstrated in this and earlier works is that following the finite there is a *transfinite* (which one could also call the *supra-finite*), that is an unbounded ascending ladder of definite modes, which by their nature are not finite but infinite, but which just like the finite can be determined by definite well-defined and distinguishable *numbers*. [Hal84, p. 39]

Cantor defined ordinal but not cardinal numbers in the *Grundlagen* (cardinal numbers came later in the same year), and so it is ordinal numbers that Cantor takes to be basic in this passage.

Since ordinal numbers play such a central role for Cantorian set theory, it is worthwhile to see how Cantor conceived of them. The ordinal numbers are, according to Cantor, generated by two principles: each ordinal number has an immediate successor, and each unending sequence of increasing ordinal numbers has an ordinal number as its limit (that is, there is an ordinal that is the next after such a sequence). He made this precise using the notion of a well-ordered collection: a collection M is *well-ordered* by a relation $<$ if $<$ linearly orders M with a least element and every subset of M that has an upper bound not in it has an immediate successor.[11] Two well-ordered sets are of the same type if they "can be related to one another one to one and uniquely in such a way that the *sequence of elements is reciprocally preserved.*" Finally, an (ordinal) number is defined to be "the symbol or concept for a definite type of well-ordered set." (See [Hal84, pp. 49–52].)

[10] In this section, I have relied heavily on Michael Hallett's illuminating work [Hal84], from which most of the translations of Cantor's words are taken. My own analysis of Cantor's work, which in many respects takes Hallett's as a starting point, is presented in §IV.2.

It is reasonably clear that, on the one hand, the two principles represent an attempt to characterize the process that generated Cantor's transfinite symbols, and that, on the other, the notion of a well-ordering isolates key features of the sequence generated using the two principles. In the *Grundlagen*, Cantor regarded the process of bringing a set into the form of a well-ordered set, thereby specifying a definite succession of the elements of the set, as giving a way of *counting* the members of the set [Hal84, p. 146]. It is not hard to see why, though Cantor did not, so far as I know, explain in detail: When one counts a set, the order in which the members are counted linearly orders the set in such a way that there is a first member, each member has an immediate successor, and there is a next member after every sequence of members counted that has not exhausted the set [Hal84, p. 63]. That is, counting a set produces a well-ordering of it.[12] Conversely, if one has a well-ordering of a set, one can count it by following that well-ordering: one counts off the first element of the set (in the sense of the well-ordering) first, the successor of each element after that element, and if a sequence of members of the set has not exhausted the set and therefore has an upper bound not in the sequence, one counts the immediate successor of the sequence next.

Cantor clearly affirmed his commitment to the equivalence between counting and well-ordering when he said (in the *Grundlagen*)

> that definite countings can be effected both on finite and on infinite sets, assuming that one gives a definite law according to which they become *well-ordered* sets. That without such a lawlike succession of the

[11] A member m of M is an upper bound of a subset N of M if m is not less than any member of N. A member m of M is a least upper bound of N if m is an upper bound of N such that if l is any other upper bound of N, then m is less than l. A member m of M is an immediate successor of N if m is an upper bound of N not in N such that if l is any other upper bound of N that is not in N, then m is less than l.

[12] I am not here endorsing Cantor's view that one can make sense of a notion of counting for infinite sets, merely describing how it seems to me the view must go.

> elements of a set no counting with it can be effected lies in the nature of the concept *counting*. [Hal84, p. 146]

The sense in which the ordinal numbers are basic is now easily stated: The transfinite sets are those that can be counted, or, equivalently given Cantor's analysis of counting, those that can be numbered by an ordinal or well-ordered. The extent to which that idea is central to Cantor's thinking is indicated by the fact that in the *Grundlagen* he refers to all sets as "countable" [Hal84, p. 150].[13] The sets are the well-orderable, increasable manifolds or classes. Unincreasable manifolds are *not* sets.

> **Technical Remark.** Cantor was well aware that the same infinite set can be counted in various ways. Indeed, as Hallett said [Hal84, p. 151],"Cantor remarks [in the *Grundlagen*] that the *only* essential difference between finite and infinite sets is that the latter can be enumerated (counted) in various ways while the former can be enumerated in just *one* way." Cantor associated a *number class* with each infinite (ordinal) number γ: the class of all numbers of the same power as γ. He then used the number classes to represent the various powers, which did not become associated with a new sort of number, a cardinal number, until later. As discussed in §2, he knew that each number class has a power greater than that of any of its members. That formed part of an argument that each power "is coordinated" with a number [Hal84, p. 41]. Thus Cantor's theory of powers was based on his theory of ordinal numbers [Hal84, pp. 62, 65]. We shall have no need to discuss Cantor's theory of powers. The interested reader would do well to consult [Hal84].

Cantor's heuristic provides an immediate counterexample to the idea that his theory can be a theory of all classes: the class of ordinal numbers obviously cannot be counted, that is, assigned an ordinal number, since any ordinal number has many successors by Cantor's principles of generation. Briefly, the class of all numbers cannot be numbered. As Cantor himself put it in the *Grundlagen*,

> I . . . now consider the only problem to be to investigate the relations of these supra-finite numbers, not just mathematically but also quite generally in tracking down and demonstrating where they appear in nature. I

[13] "Countable" has come to mean finite or denumerable. I avoid that usage because of the clash with the terminology of the *Grundlagen*.

> have no doubt at all that in this way we extend ever further, never reaching an insuperable barrier, but also never reaching any even approximate comprehension of the Absolute. The Absolute can only be recognized, never known, not even approximately. For just as . . . given any finite number, no matter how large, the power of the finite numbers following it is *always* the same, so following each supra-finite number . . . there is a totality of numbers . . . which has lost nothing as regards power with respect to the whole of the absolutely infinite totality of numbers beginning with 1 . . . The absolutely infinite sequence of numbers therefore seems to me in a certain sense a suitable symbol of the Absolute. [Hal84, p. 42]

Cantor's procedure was the reverse of what one might expect, and that is the source of much of his success, as Hallett convincingly argued [Hal84]. To obtain a theory of number that applies to both the finite and the infinite, Cantor restricted himself to those infinites that are like the finite in that they can be counted. Of course, we cannot actually count infinite sets. Frege pointed out that we can't even count sufficiently large finite sets, and so Cantor was using an idealization or extension of the usual notion of counting. Most attempts at giving a rationale for Cantorian set theory run head-on into the problem of justifying a suitable idealization. That was not a problem for Cantor, whose position was straightforwardly theological: for him "countable" meant countable by God [Hal84, pp. 15, 35–36, 44].

Frege, in contrast, took the expected approach: since infinite collections *cannot* be counted, he sought a theory of number that is independent of counting. He therefore took one-to-one correspondences to be basic, not well-orderings. That resulted in a theory in which the cardinal numbers are basic, not the ordinal numbers. (See [Hal84, pp. 151–153] for a detailed discussion.) No restriction seems necessary: a Fregean theory will not obviously need to exclude "the Absolute."

Cantor's argument that "the absolutely infinite sequence of numbers" is "a suitable symbol of the Absolute" can be turned into a mathematically precise proof that the class of all ordinal numbers is not a set. Cantor gave such a proof in a letter to Dedekind dated 3 August 1899 [Can32a].[14] According to a letter he wrote to Jourdain (dated 4 November 1903, see [GG71, p. 117]), Cantor knew that proof by 1895. Since he could now obtain a contradiction

[14] See [GG74, pp. 127–128] for the reason the letter is dated incorrectly in [vH67].

from the assumption that the system of all numbers is a set, he began to call it an *inconsistent* absolutely infinite multiplicity.

Technical Remark. Here is the proof. (I have departed inessentially from Cantor's version.) As we have seen, to every well-ordered set F corresponds an ordinal number, $\overline{F}$. Thus, to use our earlier examples, $F = (a_1, a_2, \ldots)$ has ordinal number ω, that is, $\overline{F} = \omega$, while $G = (b_2, b_3, \ldots, b_1)$ has ordinal number $\omega + 1$. There is a natural sense in which ω is less than $\omega + 1$: there is an initial segment of G, namely the part of G that comes before b_1, that has order type ω, as the one-to-one correspondence that pairs off a_i with b_{i+1} shows.

As Cantor had proved, the ordinal numbers are well-ordered by the following natural order, which generalizes what was just illustrated for ω and $\omega + 1$: If α and β are both ordinal numbers, we say that α is less than β if there is a well-ordered set F such that $\overline{F} = \beta$ and a member a of F such that *the initial segment of F determined by a* has order type α, where the initial segment of F determined by a is just the subset of F that consists of members of F less than a, with the same order they had in F.

The natural order on the ordinal numbers has the following convenient property: For any ordinal α, the set of ordinal numbers less than α ordered by the natural order form a well-ordered set of order type α. Thus, for example, the set of ordinal numbers less than 3 ordered in the usual way, that is, $(0, 1, 2)$, is a well-ordered set of order type 3.

Let Ω be the class of all ordinal numbers, and suppose that Ω is a set. Then Ω is a set well ordered by the natural order, and so it has a corresponding ordinal number, say $\overline{\Omega} = \delta$. Thus, Ω is a well-ordered set of type δ. But, by the definition of Ω, the ordinal number δ must be a member of Ω. By the convenient property mentioned above, the initial segment of Ω determined by δ has order type δ, and so, by the definition of the natural order, δ is less than $\overline{\Omega}$. That is, δ is less than δ, which is impossible. The contradiction shows that our initial assumption that Ω is a set must be false. As Cantor put it [Can32a, p. 115], "The system Ω of all numbers is an inconsistent, absolutely infinite multiplicity."

Since Cantor had already argued that to every ordinal number corresponds a distinct cardinal number,[15] it followed from the fact that the system of all ordinal numbers is an inconsistent, absolutely infinite multiplicity that the system of all the cardinal numbers that correspond to ordinal numbers is also an inconsistent, absolutely infinite multiplicity.[16] Cantor went on to employ those results to prove certain theorems he had been working on for a long time. The details belong in §IV.2.

§5. Paradoxes

In 1895, the year in which Cantor discovered the arguments just presented, Bertrand Russell wrote his fellowship dissertation, which was to become *An Essay on the Foundations of Geometry* (1897). It was a neo-Hegelian work: Russell believed that every science (except the universal science, metaphysics) necessarily contains contradictions that require a dialectical transition to another science for their resolution. For example, geometry is the science of pure spatial relations. But relations need something to relate. Thus geometry must postulate something beyond pure spatial relations: spatial points. The contradiction is transcended by moving toward physics. (See [Gri88, pp. 20, 24–26].)

In 1896, Russell learned of Cantor's work. Russell later said [Rus67a, p. 200], "At that time I falsely supposed all his arguments to be fallacious, but I nevertheless went through them all in the minutest detail. This stood me in good stead when later on I discovered that all the fallacies were mine." At the time, Russell believed that [Gri88, p. 32] "the continuum as an object of thought is self-contradictory." In 1897, Russell reaffirmed his doubts about the mathematical infinite, but in 1898 he tentatively accepted it, in an early draft of what was to become *Principles of Mathematics* [Rus03] (see [Moo88b, pp. 49, 50]).

Russell's acceptance of the infinite did not last long. In 1899, around the time Cantor was writing the letter to Dedekind discussed in the previous section, Russell was lecturing and writing about Leibniz. Leibniz accepted the actual infinite but argued against infinite number. In the second draft of what was to become the *Principles,* written in 1899, Russell accepted that infinite number is contradictory but worried that a class, the extension of a concept (that is, the collection of things to which the concept applies),

[15] The number classes inherit their ordering from the ordinal numbers, and they are therefore well-ordered. Since every sequence of cardinal numbers has an upper bound, they are absolutely limitless, and so they must form a well-ordered class similar to Ω.

[16] The argument bears some relation to the later Axiom of Replacement: the range of a function on a set is a set. See §V.2.

is a totality, which should therefore have a number. The simplest version of the contradiction arises, he observed, when one considers the totality of numbers. He quoted Leibniz as saying, "the number of all numbers implies a contradiction," and he wrote: "There is, and is not, a number of numbers." (See [Moo88b, p. 50] and [MG81, p. 325].) The problem whether infinite collections have infinite number or no number continued to dog him in the subsequent draft, which he worked on into the next year.

In the summer of 1900, Russell met Peano at a conference and was favorably impressed. He began to study Peano's work. He later said,

> The time was one of intellectual intoxication. My sensations resembled those one has after climbing a mountain in a mist, when, on reaching the summit, the mist suddenly clears, and the country becomes visible for forty miles in every direction. For years I had been endeavouring to analyse the fundamental notions of mathematics, such as order and cardinal numbers. Suddenly, in the space of a few weeks, I discovered what appeared to be definitive answers to the problems which had baffled me for years . . . Intellectually, the month of September 1900 was the highest point of my life. [Rus67a, pp. 232–233]

As a result of studying Peano's work, apparently in September, Russell came to accept that every collection has a cardinal number. By November, he had found an "error" [Cof79, p. 33] in Cantor: Cantor's diagonal argument proved that there is no largest cardinal number. But the number of individuals is the largest number, since every class is included in the class of individuals. (Russell counted classes and numbers as individuals.) At around the same time he also noted that if the ordinal numbers are, as Cantor claimed, well ordered, then there is a maximum ordinal number, namely, the order type of the class of all ordinal numbers. He also described the error as involving the class of classes instead of the class of individuals.[17] (See [Moo88b, pp. 52–53].)

Russell did not yet suspect any paradox, though he had found a contradiction. He believed that Cantor's conclusions were not quite so general as they seemed. In detail, he doubted Cantor's assertion that the ordinal numbers are well ordered, and he supposed that Cantor's diagonal argument, which he took to show that the class of all subclasses of a class is of strictly greater power (cardinal number) than the class is not quite so general as it seemed—it does not apply to the class of all individuals. The last supposition is one Cantor later endorsed:[18] Cantor viewed his results as applying only to "countable" *sets*—a notion we shall discuss in detail in §IV.2—not to arbitrary collections. Russell's work concerned classes, the notion introduced here and discussed in more detail in §IV.1. But during the period under discussion—and usually even later—Russell just interpreted Cantor's work as if it concerned Russellian classes. When Cantor saw Russell's later work, he concluded that the class of all individuals was not a set at all but an inconsistent, absolutely infinite multiplicity.

Russell's suspicion that Cantor's argument did not apply to the class of classes was based on the following, quite reasonable grounds. The class of classes has as members all classes, including those that have individuals other than classes as members. The class of all sub*classes* of the class of classes, in contrast, is composed entirely of classes that have only classes as members. It must therefore be a proper part of the class of classes, and so it cannot be of larger power than the class of classes. (See [Cof79, p. 34].)

Russell took Cantor's argument to show that the class of all subclasses of a class has greater power than the class, and he recast the argument in essentially the following way:[19]

(1) He first showed for any function k from a class u to the class of all subclasses of the class u, that the class of all members x of u such that x is not in k_x is a subclass of u not in the range of the function k.

(2) He then observed that the class of subclasses of a given class has power at least as great as that of the class, since the function that takes each member x of the given class to the class whose only member is x is a one-to-one correspondence between the given class and some of its

[17] I do not know why Russell shifted from the class of individuals to the class of classes. It is not hard to argue that the two classes have the same cardinal number, and conclude that if the cardinal number of one of them is the largest cardinal number, then so is the cardinal number of the other. One argument, which Russell definitely gave later [Rus03, p. 367], goes like this: The class of classes is contained in the class of individuals, and so it is no larger. Conversely, the class of classes with exactly one member is the same size as the class of individuals (since each individual corresponds with the class that has it and nothing else as a member), and the class of classes with exactly one member is contained in the class of classes, and so it is no larger. Neither class is larger than the other, and so they have the same cardinal number, as required.

[18] In a letter translated in part as Appendix A to Chapter IV.

[19] Cantor's point of view was different. His paper is translated as Appendix B to Chapter IV.

subclasses. To show that the class of subclasses of a class has greater power than the class, it therefore suffices to show that the two do not have equal power.

(3) Finally, he supposed for the sake of contradiction that a class and the class of its subclasses have equal power. Then there is a function from the class to the class of its subclasses that establishes a one-to-one correspondence between them. But that contradicts item 1: a one-to-one correspondence cannot omit a member of the range.

Russell thought the argument was in error when the "class" was the class of classes. Following Russell, let *class* be the class of classes. We can, it seems, define a function k from *class* to the class of subclasses of *class* that includes every subclass of *class* in its range as follows: when x is in *class* and x is a class of classes, let k_x be x, and when x is in *class* and x is not a class of classes, let k_x be the class whose only member is x. But that violates item 1 of the Cantorian proof: According to the proof of item 1, the class u' of classes x such that x is not a member of k_x should not be in the range of k. But, Russell observed, u' is $k_{u'}$ (presumably since k is the identity function on classes of classes), and so, contrary to the Cantorian argument, u' *is* in the range of k. Thus, Russell concluded, item 1 of Cantor's argument is incorrect when the class involved is *class* and the function is k, and so Cantor had not shown that there is no largest cardinal number. Russell gave this analysis by November of 1900. He added, seemingly as an afterthought, that "in fact, the procedure is, in this case, impossible; for if we apply it to u' itself, we find that u' is a $k_{u'}$, and therefore not a u'; but from the definition, u' should be a u'." (See [Cof79, pp. 35–36].)

Russell seems to have maintained the view that Cantor's argument is defective and that there is a largest cardinal number at least through the middle of January 1901 [Cof79, p. 33]. But the u' of the above argument is readily seen to be the class of classes that do not belong to themselves, and the above afterthought just shows that u' both is and is not a member of itself. That is, the definition of u' leads to a contradiction. There is no class of all classes that do not belong to themselves.[20] Russell discovered that[21] by May [Moo88b,

p. 53]. There is no class u', and so Russell had not produced a counterexample to Cantor's argument. In October 1901, Russell wrote to Louis Couturat that Cantor is irrefutable [Cof79, p. 37].

Russell did not know what to make of his contradiction:

> It seemed unworthy of a grown man to spend his time on such trivialities, but what was I to do? There was something wrong, since such contradictions were unavoidable on ordinary premisses. Trivial or not, the matter was a challenge. Throughout the latter half of 1901 I supposed the solution would be easy, but by the end of that time I had concluded that it was a big job. [Rus67a, p. 236]

Russell finally wrote to Peano about "the matter" and to Frege in June 1902. The letter to Frege [Rus67b] introduced the argument with some diffidence: "There is just one point where I have encountered a difficulty." But Frege's attitude was clear. He replied [Fre67], "not only the foundations of my arithmetic, but also the sole possible foundations of arithmetic, seem to vanish." That is how Russell's "conundrum" became Russell's paradox. By September, the paradox was a central problem for Russell. (See [MG81, p. 328].)

Russell had uncovered two more paradoxes by the time the *Principles* appeared:[22] the paradox of the largest ordinal and the paradox of the largest cardinal. The paradox of the largest ordinal is this: The class of all ordinal numbers is apparently well ordered, and so it has an ordinal number as order type, which must be the largest ordinal. But there cannot be a largest ordinal number since every ordinal number can be increased by 1. (See [Rus03, p. 323].) The similarity between that argument and the one Cantor used to show that the ordinal numbers form an inconsistent multiplicity should be clear. The paradox of the largest ordinal has come to be known as Burali-Forti's paradox, since Russell attributed it to Cesare Burali-Forti. In fact, the paradox is due to Russell, though it was apparently suggested to him by his reading of a paper by Burali-Forti [MG81].

The paradox of the largest cardinal is this: The class of classes can be no

[20] Zermelo discovered the paradox independently, but little is known about the details. See [RT81].

[21] Actually, so far as I know, he did not at the time discuss the class of all classes that are not members of themselves, but only the class of all predicates that cannot be predicated of themselves. The class version appears in his letter to Frege [Rus67b], which he wrote a year later.

[22] My method of counting paradoxes is somewhat arbitrary. For example, I am counting the paradoxes of the class of classes that are not members of themselves and of the class of predicates that are not predicable of themselves as one, because of their evident similarity.

larger than the class of individuals, since it is contained in the class of individuals. But the class of classes is the class of all subclasses of the class of individuals, and so Cantor's diagonal argument shows it to be larger than the class of individuals. Russell introduced that paradox as follows [Rus03, pp. 366–367]: "[Cantor's] argument, it must be confessed, appears to contain no dubitable assumption. Yet there are certain cases in which the conclusion seems plainly false." The paradox is often called Cantor's paradox, presumably because it is based on Cantor's argument. It may be rephrased as follows: Cantor's argument shows that there is no largest cardinal number. But the cardinality of the class of all individuals must be the largest cardinal number, since every other class is included in that one.

Russell's paradox seems the most important, because it is so much more direct than the others. The paradox of the largest cardinal in some sense already involves Russell's paradox, as we have seen from the manner in which Russell discovered his paradox. The paradox of the largest ordinal involves the machinery of well-ordered sets and ordinal numbers, and so Russell thought it might be dissolved in some technical way.

IV

What Are Sets?

§1. Russell

Why did Russell find paradoxes where Cantor found none? Because Russell accepted a principle that Cantor did not, one that conflicted with principles on which they agreed. The extra principle, which seems to have originated with Peano, was the Comprehension Principle. In Russell's words [Rus03, p. 20], "a class may be defined as all the terms satisfying some propositional function." In that and most other respects concerning the notion of class Russell followed Peano, as he said quite clearly. In particular, for Russell, a class is "composed of terms."

The Peano–Russell notion of class is essentially what Penelope Maddy has called the logical notion of collection. The characteristic mark of the notion is that according to it each collection is associated with some kind of a definition or rule that characterizes the members of the collection.[1]

Frege had a notion in the *Grundgesetze der Arithmetik* that is formally equivalent to that of a class, and a principle analogous to the Comprehension Principle—a principle that subjects his system to the paradoxes. Nonetheless, as Russell noted [Rus03, p. 513], Frege did not countenance classes in the now familiar form that comes down to us from Peano via Russell.[2]

[1] The term is Maddy's, but she used it slightly differently [Mad90, pp. 103, 121]: "The logical notion . . . takes a number of different forms depending on exactly what sort of entity provides the principle of selection, but all these have in common the idea of dividing absolutely everything into two groups according to some sort of rule." Compare [Göd47, p. 475].

[2] Frege took concepts to be basic. He was interested in a particularly important equivalence relation between concepts, that of extensional equivalence: two concepts are extensionally equivalent if they hold of the same objects. He postulated that to each con-

Russell was at least dimly aware that Cantor's conception of a set was different from his own:

> when mathematicians deal with what they call a manifold, aggregate, *Menge* [Cantor's term], *ensemble,* or some equivalent name, it is common, especially where the number of terms involved is finite, to regard the object in question (which is in fact a class) as defined by the enumeration of its terms . . . Here it is not predicates and denoting that are relevant, but terms connected by the word *and,* in the sense in which this word stands for a *numerical* conjunction. [Rus03, p. 67]

Cantor's conception, which is discussed in detail in the next section, forms the basis for the one almost universally used by mathematicians. The main evidence for that claim is presented in §V.1. After that point, we give a detailed history of the Cantorian conception with practically no need to refer to

cept corresponds a logical object, the *extension* of the concept, in such a way that extensionally equivalent concepts correspond to the same object, while concepts that are not equivalent do not.

Frege did not have much more to say about the nature of his logical objects—his "extensions." They are often confused with Peano's classes because propositional functions that are satisfied by the same objects determine the same class—a property formally analogous to the one postulated by Frege. They are not the same: classes are composed of terms, and so the membership relation was basic for Peano, but Frege's logical objects were defined without reference to membership. To be sure, Frege later defined a notion formally equivalent to membership as follows: x is a "member" of the logical object y if there is some concept F such that y is the logical object that corresponds to F and x falls under the concept F. But that was clearly not the basis for his logical objects. Indeed, Frege said [Fre95, p. 228], "the concept is logically prior to its extension; and I regard as futile the attempt to take the extension of a concept as a class, and make it rest, not on the concept, but on single things." He summed up: "The extension of a concept does not consist of objects falling under the concept, in the way, e.g., that a wood consists of trees; it attaches to the concept and to this alone. The concept thus takes logical precedence of its extension."

Moreover, Frege wanted everything in his system to be one of his logical objects, so he just arbitrarily stipulated that the object that is the extension of the concept "x is the true" is the true, *not* the class of truths, and, similarly, that the object that is the extension of another concept is the false. That would not have been possible if he had intended the Peano–Russell notion of a class composed of members. (See [Rus03, pp. 510–512] and [Res80, pp. 204–220].)

I am under the impression that as late as 1903 Frege did not fully understand the Peano–Russell notion: Frege attempted to "avoid the contradiction" (Russell's paradox) by permitting two concepts to correspond to the same object even though that object "falls

the Peano–Russell conception. That in itself will help to show that the Cantorian conception is dominant. But one can also find an explicit statement of that dominance in an article by Skolem to be discussed in detail in §V.3:

> Until now, so far as I know, only *one* such system of axioms [for set theory] has found rather general acceptance, namely, that constructed by Zermelo [Zer08b]. Russell and Whitehead, too, constructed a system of logic that provides a foundation for set theory; if I am not mistaken, however, mathematicians have taken but little interest in it. [Sko23b, p. 291]

As we shall see in §V.1, Zermelo's system is an outgrowth of Cantor's. I do not wish to leave the impression that Frege and Russell are or were unimportant. It is only that their mathematical work was for the most part concerned with logic, formalization, axiomatization, and related issues, not with the theory of sets. In particular, the paradoxes are important for what they tell us about our conceptions of properties and of truth, but they are not important for the theory of sets, as Gödel had already observed in 1947 [Par86, p. 105].

It is necessary to discuss Russell's conception of classes as it developed in response to the paradoxes before turning to Cantor's conception of sets, if only to make sure that the two are clearly distinguished. Russell was a logicist. He wished to show that mathematics and logic are one by showing how to develop all of mathematics within a framework free of any special

under the one and not under the other." (See [Fre80, p. 150].) (Frege called that object the extension and Russell used the term *range of values*, which was his translation of a term of Frege's that includes but is more general than Frege's *extension.* Since what is at issue is whether their use of the terms is similar to ours, I have been avoiding those terms.) Russell queried [Fre80, p. 155], "Do you believe that the range of values remains unchanged if some subclass of the class is assigned to it as a new member?" Frege replied [Fre80, p. 157], "I do not believe that a class remains in general unchanged when a particular subclass is added to it. All I mean is that two concepts have the same extension (the same class) when the only difference between them is that this class falls under the first concept but not under the second." Whatever Frege's extensions are, their members are not constitutive of them. The fact that he identified them with classes in the passage shows that he had not understood the notion of class. According to Charles Parsons, Frege always took classes to be either (Fregean) extensions or aggregates made up of parts, that is, mereological sums [Par76, p. 268].

conditions or empirical and psychological assumptions. That is a program substantially similar to Frege's for arithmetic and analysis.[3]

Frege and Russell faced a common problem: mathematics is apparently about objects (numbers and so forth), and yet the assumption that objects of one sort or another exist apparently goes beyond pure logic. They arrived at formally analogous solutions, which I explain with an example. Definable relations *are* within the province of logic, including definable one-to-one correspondences, and so, without going beyond logic, one can use Cantor's method to define *equinumerosity:* two systems[4] are equinumerous if there is a one-to-one correspondence between them. Suppose one postulates as a *logical* principle that every equivalence relation (equinumerosity in the example) determines logical objects and a logical relation such that entities in the field of the equivalence bear the relation to the same logical object if and only if they are equivalent. Then the logical objects will be suitable to play the role of mathematical objects. Russell used classes as the logical objects and membership as the logical relation (see [Rus03, pp. 166–167]), as was suggested by the work of Peano, Burali-Forti, and Mario Pieri (see [Con87]). Thus, for Russell a number was a class of all systems equinumerous to any member of the class. For example, on Russell's account, the number 2 is the class of all pairs. Thus, to be a system of two objects is just to be a member of the number 2, that is, a member of the class of all pairs. In general, the number of a system was simply the number of which the system was a member. For Frege, the number of a system was the extension of the concept "being equinumerous with that system."

Russell was ≡ [illegible] [illegible] but C. used [illegible] ϕ

The Comprehension Principle was what provided the mathematical objects on Russell's early logicist account of mathematics. It played a central role. But the Comprehension Principle was the source of the paradoxes. Russell therefore had two options: restrict the propositional functions to which the principle applied or restrict the propositional functions themselves, so that the principle still held outright. Whatever restrictions he adopted had to be purely logical in character. He tried both options, and gave many variants on the second one. I shall only discuss one of his theories, which is of the second kind: his 1908 theory of types, as presented in [Rus08]. I have focused on that theory since I understand it the best, and since it raises some points that will be useful in subsequent chapters.[5]

[3] Russell's motivation was rather different from that of Frege. See [Hyl90].

[4] I am using *system* here as a neutral word for whatever has an associated number. Frege and Russell disagreed.

$Het^2(P) \Leftrightarrow P'(P)$ v violates ST?

In the 1908 theory of types, individuals and propositions are taken to be basic,[6] and sentences that apparently mention propositional functions and classes are analyzed as involving only basic entities. In that sense, the theory is a no-class theory—propositional functions and classes are not taken to be "part of the ultimate furniture of the world."

I shall discuss propositional functions in some detail before getting to classes, since most of the work is done by the theory of propositional functions. The propositional function "x is mortal," for example, can be represented by the pair of the proposition "Socrates is mortal" and the individual Socrates. Russell took as primitive the quaternary relation exemplified by: the result of substituting *Plato* for *Socrates* in *"Socrates is mortal"* is *"Plato is mortal."* That has the following advantage: If we take propositional functions as basic, we run straight into paradoxes, as is seen by substituting the propositional function "x is not self-predicable" into itself. However, if we adopt Russell's stratagem, then "x is not self-predicable" must seemingly be represented by some pair like "Socrates is not self-predicable" and Socrates or " 'Socrates is mortal' is not self-predicable" and "Socrates is mortal," depending on whether we take the variable in "x is self-predicable" to range over individuals or propositions. Clearly, neither captures our intent—the variable was supposed to range over propositional functions. The one variable must therefore become two: "(p, a) is self-predicable," where now 'self-predicable' must mean something concerning a pair. But "(p, a) is self-predicable" is a propositional function that is to be represented (since there are two free variables) by a triple: "('Socrates is mortal,' Socrates) is self-predicable," "Socrates is mortal," Socrates. We needn't worry about how to make sense of 'self-predicable' because there is a different kind of trouble—'self-predicable' was to be defined for propositional functions represented by pairs, but the propositional function we intended to substitute to obtain a para-

[5] The article [Urq88] provides a useful brief history of Russell's attempts to solve the paradoxes. The articles [Lan87] and [Lan89] are very helpful in understanding the development of Russell's theory of types, and I have relied on them heavily. ✓

[6] According to Peter Hylton [Hyl90, pp. 151, 155], Russell had just postulated classes, but he had an argument for the existence of individuals and propositions, namely that they are required for logic.

TT an FOL with vbls $v_j^{\tau_0}$ s.t. $\tau \in T = \bigcap\{x \mid \tau_0 \in x \vee \forall \tau_1,\dots,\tau_n \in x(\langle\tau_1,\dots,\tau_n$

τ_0: type of ind

$\langle\tau_0\rangle$: [illegible]

dox is represented by a triple, and so it cannot be substituted. The paradox is blocked!

The analysis of propositional functions into propositions and individuals creates in effect a hierarchy of types, with propositions and individuals at the bottom, with single variables ranging over them; propositional functions of propositions or individuals next, with pairs of variables ranging over them; (3) propositional functions of (propositional functions of propositions or individuals next) and so forth. (The types are not in strict linear order: there can be propositional functions of both individuals and propositional functions of individuals, and so forth. See [Rus08] or [Lan87] for details.) According to Henri Poincaré, a vicious circle of definitions is the source of the paradoxes. Here, no propositional function can have itself in its own range, and an analogous circle is blocked.

Note the great ingenuity of Russell's device. It falls out of Russell's system of representation that the old notion of propositional function was incoherently wide. Eliminating the incoherence, of course, has the effect of restricting which propositional functions are allowed. But the restriction, while it suffices to block the paradoxes, allows one to retain the air of perfect generality: eliminating the use of incoherent propositions, while it is the required restriction on what came before, is not a defect in logical purity, and when one begins with the new system of representations, it need not be presented as a restriction. All the usual devices of logic can be allowed, and the types arise without any special pleading. The theory can still lay claim to being a part of pure logic, and so mathematics might still be one with logic, if no additional modifications were required.

Unfortunately, the typed system just described is still subject to paradox, as Russell had realized by 1906 [Lan89, p. 37].[7] The present theory allows quantification over all propositions and individuals, and hence, derivatively, over all propositional functions of a single type, including propositional functions of a type that are specified using quantification over propositional functions of that very same type.

Technical Remark. The paradox Russell discovered involved propositional functions of propositions. He worked directly in his basic sys-

tem, in which quantification over pairs of propositions stands in stead of quantification over such propositional functions. I shall use quantification over propositional functions instead, since the argument becomes easier to follow in that notation, but such quantification is just an abbreviation for something more complicated in the base notation. The simplification is one Russell had adopted by 1908. All of my Latin variables will range over propositions and Latin constant symbols stand for propositions, while my Greek variables range over propositional functions and Greek constant symbols stand for propositional functions. Since we shall explicitly consider only propositional functions of one variable, we can use, for example, ϕ to indicate a propositional function and $\phi(x)$ to indicate the value of that function at x. Let ψ be the propositional function (of y) *

$$\psi(y) = \quad (\exists\phi)(y = [\phi(b) = q] \wedge \neg\phi(y)).$$

I have respected modern scruples about use and mention to the extent of forming a name for a proposition by enclosing the proposition in square brackets. Russell had no such scruples.[8] Now consider the proposition $\psi([\psi(b) = q])$, which reads as follows:

$$(\exists\phi)([\psi(b) = q] = [\phi(b) = q] \wedge \neg\phi([\psi(b) = q])).$$

The equality $[\psi(b) = q] = [\phi(b) = q]$ is between propositions, which are intensional. Thus, when the equality holds, it follows that $(\forall y)([\psi(y)] = [\phi(y)])$ and hence that $(\forall y)(\psi(y) \leftrightarrow \phi(y))$. We can therefore derive the contradiction $\psi([\psi(b) = q]) \leftrightarrow \neg\psi([\psi(b) = q])$.

Once more we have a circle of substitutions—ψ has been substituted into itself.

To block the new paradox, Russell introduced "orders" of propositions. At the bottom, there are the first-order propositions: "elementary" propositions and those that involve quantification only over individuals. Next come

[7] I am simplifying the story somewhat. In 1906, Russell was working with a type theory of complexity intermediate between that of the one described in the text and the 1908 version toward which I am working.

[8] Since Russell was happy to allow truth predicates in his base notation, it is possible to reformulate the paradox in a way that meets modern standards, though I shall not stop to do so here.

[illegible] has [illegible]: $\forall \alpha \in E$, ${}^{m}E$ iff $m = \max_x$ [illegible] $V \in V$ is contained in [illegible]

[illegible] del [illegible] ${}^{m}E^{n} = {}^{M}E^{n}$

def. $[\forall y.] \psi(y) \Leftrightarrow \exists \phi (y = [\phi(b) = q] \wedge \neg\phi(y))$ Example of a semantic parad

$\psi[\psi(b) = q] \Leftrightarrow \exists\phi([\psi(b) = q] = [\phi(b) = q] \wedge \neg\phi[\psi(b) = q])$

second-order propositions, which may also involve quantification over first-order propositions. And so forth. Quantification is not permitted over all propositions, but only over propositions of a given order.

We have now described Russell's "ramified" hierarchy. Propositional functions of any given type may be defined by propositional functions of various orders. The system, while less natural than before, is still motivated exclusively by logical concerns. After all, paradoxes cannot be permitted, and so the argument above might be taken to show that orders of propositions (or of whatever may serve as a surrogate for propositions) are logically necessary. The ramified hierarchy is indeed Russell's proposed logical system, though he modified it in various ways subsequent to 1908. Unfortunately, additional assumptions are required to do mathematics. I shall explain after showing how Russell handled classes.

Propositional functions may be intensional. For example, someone (call her Caila) might believe that all humans are mortal without believing that all featherless bipeds are, despite the fact that "x is human" and "x is a featherless biped" are coextensive.[9] Thus, the propositional function "Caila believes that for all x if $\phi(x)$ then x is mortal" depends for its truth value on the particular function ϕ, not just on what satisfies ϕ. Say that a propositional function Θ is extensional if

$$(\forall\phi\psi)((\forall x)(\phi(x) \leftrightarrow \psi(x)) \rightarrow (\Theta(\phi) \leftrightarrow \Theta(\psi))),$$

that is, if its truth value is the same on coextensive propositional functions ϕ and ψ.[10] For extensional propositional functions used only within extensional propositional functions, which are all we need for mathematical purposes, we can simply identify each function with the class of things that satisfy it and then define, for example, $\phi \in \Theta$ to be $\Theta(\phi)$, $\phi \cap \psi$ to be $\phi \wedge \psi$, and so forth. We make analogous definitions for propositional functions of individuals, and also for relations and functions as well as classes.

But now we are in trouble, because relations, classes, and functions, since after all they are just certain propositional functions, have orders. Thus, for example, a finite class is one such that there is no function whatsoever that maps it one-to-one to a proper subclass, but we cannot use that fact to define

[9] I shall follow tradition in ignoring the existence of plucked chickens and other counterexamples to the supposed coextensiveness.

[10] All of the lower-case Greek variables here range over propositional functions of some fixed order, though it makes no difference which order. Similar remarks apply below.

RTST For with $^m V^n$, type n, order n (see p 72)

the finite classes, because we shall only be able to quantify over functions of some order or other. We cannot express "no function whatsoever," only "no function of such and such an order." We are faced with the possibility of a class being finite with respect to functions of some order, but infinite with respect to functions of some higher order, and the possibility of doing ordinary mathematics has apparently been lost. Russell's logicist program failed after all as a result of the paradoxes—his ingenious repair was unsuccessful.

Russell did not give up so easily as I have just made it seem. What is missing from his theory so far is a notion of class that does not have an associated order. The paradox that led us to introduce orders was an intensional one, while classes are extensional. That suggests a way out: We cannot assume that all propositional functions (of a given type) are of the same order, because of the paradox, but we can fix an order for any given type and then assume that every propositional function (of whatever order) of the given type is coextensive with one of the fixed order. That is to say, we can assume that some order includes so many propositional functions that they can represent all the classes. Since all the classes will then be of the same order, we shall be able to quantify over all of them. The assumption that allows that, and which therefore makes it possible to do mathematics within the theory of types, is Russell's Axiom of Reducibility. Details follow.

Technical Remark. Recall that a propositional function is analyzed as a proposition (the *prototype*) plus one or more propositions or individuals that in effect indicate the argument places. Let the order of a propositional function be (the least number that is) at least the order of the prototype and greater than the orders of the arguments. (Individuals have order zero; elementary propositions, order one. The definition differs from that of [Rus08, p. 164]. It makes, for example, $(\Theta(\phi), \phi)$, where ϕ is elementary, at least second order, which brings the orders of propositional functions into line with those of [WR57, pp. 163–164].) Say that a propositional function is *predicative* if its order is one greater than the orders of the propositions and individuals indicating the argument places. Following Russell, we shall indicate that a propositional function is predicative with an exclamation point, thus: $\phi!$.[11] The Axiom of Reducibility reads

[11] Alfred North Whitehead and Russell [WR57, pp. 164–165] defined a predicative propositional function to be one that is quantifier-free, which is fine for the statement of the Axiom of Reducibility, perhaps even more natural than what I have done here.

$\exists\, {}^{1}\varphi^{T}_{n}\ \forall x_1,\ldots,x_n\ [{}^{1}\varphi^{T}_{n}(x_1,\ldots,x_n) \Leftrightarrow {}^{n}\varphi^{T}_{n}(x_1,\ldots,x_n)]$

Reducibility

$$(\exists\phi)(\forall x)(\phi!(x) \leftrightarrow \psi(x)),$$

where ψ is a schematic letter that may be replaced by any propositional function of any type and order with one free variable, and x is a variable of the appropriate type and order. The axioms for two or more free variables are analogous. Russell also refers to the axiom for one free variable as the Axiom of Classes and to the axiom for two free variables as the Axiom of (Binary) Relations.

We are not quite out of the woods yet: we cannot take the class associated with a propositional function to be the corresponding coextensive predicative propositional function, since there may be many coextensive predicative propositional functions. We therefore give a contextual definition that shows how to reinterpret formulas involving classes as formulas without them. They are to be eliminated from our official vocabulary much as propositional functions have already been eliminated.

With any propositional function Θ of a predicative argument $\phi!$, we associate a schematic formula $\Theta(\{z : \psi(z)\})$ that is defined as follows:

$$(\exists\phi)((\forall x)(\phi!(x) \leftrightarrow \psi(x)) \wedge \Theta(\phi!)),$$

where ψ is a schematic letter that may be replaced by any propositional function of a variable x appropriate to $\phi!$. For example, we define $x \in \{z : \psi(z)\}$ to be the formula that arises from our scheme using the propositional function Θ defined by $\Theta(\phi!) = \phi(x)$. Then $x \in \{z : \psi(z)\}$ is an abbreviation for

$$(\exists\phi)((\forall x)(\phi!(x) \leftrightarrow \psi(x)) \wedge \phi!(x)).$$

But they also claimed that it is possible to restrict quantification to predicative—that is, quantifier-free—propositional functions. The reason they offered is not correct, and so I have offered a definition of predicative here that makes their claim correct. In a special case, their reason amounts to this: The formula $(\forall x)\phi(x)$, where $\phi(x)$ is $(\forall y)\psi!(x, y)$, is just $(\forall x)(\forall y)\psi!(x, y)$. (In this note the exclamation point indicates that ψ is quantifier-free.) We can therefore replace $(\forall\phi)(\forall x)\phi(x)$ by $(\forall\psi)(\forall x)(\forall y)\psi!(x, y)$. But that doesn't work, since $(\forall\phi)$ includes in its range $(\forall y_1)\psi!(x, y_1)$, $(\forall y_1)(\forall y_2)\psi!(x, y_1, y_2)$, $(\forall y_1)(\forall y_2)(\forall y_3)\psi!(x, y_1, y_2, y_3)$, and so on, while the quantifier $(\forall\psi)$ that is supposed to replace it accommodates only a fixed number of y_is. Since the orders play no role in mathematical considerations after the introduction of the Axiom of Reducibility, the rest of [WR57] is unaffected.

ord. 1 ord. 2 ord. 3

Virtues of RTT

(1) RTT solves the logical paradoxes, $x \notin \{x \mid x \notin x\} \Leftrightarrow x \in \{x \mid x \notin x\}$

(2) RTT solves the semantic paradoxes,

(3) RTT is consistent (Fitch 1938, Lorenzen 1951, Schütte 1952)

As a side effect of the contextual elimination of classes, every propositional function gets associated with one that is extensional, and so we can drop the previous restriction to extensional propositional functions and contexts.

With the Axiom of Reducibility, it becomes possible to develop mathematics within the theory of types (with two notable exceptions, to be discussed just below). However, as Russell himself put it,

> Viewed from this strictly logical point of view, I do not see any reason to believe that the axiom of reducibility is logically necessary, which is what would be meant by saying that it is true in all possible worlds. The admission of this axiom into a system of logic is therefore a defect, even if the axiom is empirically true.[12] . . . There is need of further work on the theory of types, in the hope of arriving at a doctrine of classes which does not require such a dubious assumption. [Rus19, p. 193]

The earlier verdict is correct after all: Russell's logicist program failed as a result of the paradoxes.

Even if we allow the Axiom of Reducibility, there are still two gaps in the development of mathematics within the ramified theory of types: it is impossible to prove the Axiom of Infinity, which says that there is a class with infinitely many members,[13] and it is impossible to prove the Axiom of Choice (an assumption we shall be discussing in detail in §V.1). Moreover, as Russell recognized [Moo82, p. 131], the Axiom of Choice seems dubious when it is construed as an axiom concerning classes with membership specified via a rule. Those problems are serious since, for example, the Axiom of Infinity is invoked even to show that the sum of two real numbers is a real number, but they are less serious than the fundamental problem posed by the Axiom of Reducibility, since the Axioms of Infinity and Choice can simply be taken as hypotheses of every theorem in whose proof they are employed.[14] The Axiom of Reducibility must be used so universally that even the theory of the

[12] For Russell, it is an empirical matter what classes there are. Only the ones that are definable in a suitable sense exist (to the extent that classes exist at all) necessarily.

[13] Russell's axiom of infinity is that there are infinitely many individuals. It follows immediately from that in Russell's system that there is an infinite class, for example the class of individuals, and so his axiom of infinity has what we are calling the Axiom of Infinity as a consequence.

[14] Since truths about the real numbers, for example, were supposed to turn out to

Problem RTT w. Reducib.

natural numbers—let alone the rest of mathematics—would depend on it as a hypothesis, since it is required to give an adequate definition of finiteness. Indeed, since all mathematical objects turn out to be classes, and Reducibility is needed for the definition of class, not even a single mathematical object can be defined as it was defined by Whitehead and Russell without an appeal to the Axiom. To put it as starkly as possible, even the definition of the number 1 depends on the Axiom of Reducibility. That, of course, poses no problem from a technical point of view, but it hardly suffices to establish the Russellian thesis that logic and mathematics are one.

So much for the logical notion of collection. As far as the primary purpose of this section is concerned, our discussion should end forthwith. But Russell introduced a distinction that will play an important role in some of our later considerations, a distinction intimately related to the theory of types, and so we go on to present it here.

Every variable permitted in the ramified theory of types is restricted by type and order. Nonetheless, definitions, axioms, and theorems must be available at all types and orders, and so it is necessary to have devices that permit generalization of some sort across types and orders.

The first device has been much discussed. It is that of "systematic ambiguity." In actual practice we never care about the absolute types and orders of variables, but only the relative types and orders. Anything we assert about orders 1 and n holds equally about orders $1+m$ and $n+m$, and analogously for types. Thus, we can use ambiguous symbols whose relative orders and types are fixed by the context of use, and which can be applied at any absolute type and order. We normally read the variables of lowest type in a context as ranging over individuals, but, when convenient, we can reinterpret them as ranging over some higher type. Systematic ambiguity has been exploited above when we allowed Greek variables to "range over propositional functions of some fixed order, though it makes no difference which order," and when in stating the Axiom of Reducibility we took ψ to be "a schematic letter that may be replaced by any propositional function of any type and order with one free variable," and x to be "a variable of the appropriate type and order."[15] In effect the device of systematic ambiguity lets us specify schemas in which the base types and orders may take on any permissible values. For example, the theorem $(\forall\phi)\phi = \phi$ is actually infinitely many distinct theorems (one for each type and order of the variable ϕ) involving infinitely many distinct equality relations.

be logical truths, not conditional truths requiring an assumption—Infinity, Choice—that might be false, the procedure of taking such assumptions as hypotheses was not adequate to Russell's task. We take up the question whether such a procedure can be part of an adequate philosophy of mathematics of a different sort in §VI.3.

Systematic ambiguity makes it possible to express certain general facts about the formalism of ramified type theory, facts that cut across types and orders, but it does not permit us to express those facts within the formalism. After all, every quantified variable must be of a particular type and order, and so, despite appearances, $(\forall\phi)\phi = \phi$ is not a sentence expressible within ramified type theory—it is a host of separate, unconnected sentences. The trick of taking the lowest order within a formula to be that of individuals and allowing the possibility of adding m to all the orders is no help, since it is not expressible within the formalism.

To make it possible to express general facts within ramified type theory, Russell introduced the distinction between "all" and "any," which he discussed in detail in Section 2 of [Rus08]. "Given a statement containing a variable x, say '$x = x$', we may affirm that this holds in all instances, or we may affirm any one of the instances without deciding as to which instance we are affirming" (p. 156). The affirmation of $x = x$ for *all* values of x is represented $\vdash (\forall x)x = x$, and the fact that x is quantified forces us to use a variable x of fixed type and order. But the affirmation of $x = x$ for *any* value of x is represented by $\vdash x = x$. No quantifier is involved, and so we may allow a new type of variable that is free of type and order restrictions, a new type of variable that cannot be quantified over. It then becomes possible to express generality across types and orders. As Russell remarked,

> we may admit "any value" of a variable in cases where "all values" would lead to reflexive fallacies . . . the fundamental laws of logic can be stated concerning *any* proposition, though we cannot significantly say that they hold of *all* propositions. [Rus08, p. 158]

[15] Whitehead and Russell use the term *systematic ambiguity* only for cases in which all of the types and orders are determined by the choice of the base—that is, by the reinterpretation of the individual variables. It seems they may have regarded the use of a variable like ψ in the Axiom of Reducibility, which does not have its order determined by the surrounding context, as a new device. (See [WR57, p. 165].) But my extension of their terminology seems to me to be a natural and harmless one.

Russell's idea that we affirm an instance without deciding which, which he calls an ambiguous assertion, is not very clear, but it suggests taking the new variables as schematic variables that admit of substitution by variables or constants of any type and order. Some of what Russell said suggests that reading: "We can only truly assert a propositional function if, whatever value we choose, that value is true; similarly we can only truly deny it if, whatever value we choose, that value is false" (p. 157).[16]

In the first edition of the *Principia Mathematica* more or less the same distinction between "any" and "all" appeared, but at some points what I have distinguished as schematic variables are allowed and exploited [WR57, pp. 128–129], while at others, even "any" is allowed only with variables of fixed type and order [WR57, pp. 17–18]. There is no point to using "any" with variables of fixed type and order. By the second edition, the use of "any" had been repudiated [WR57, p. xiii] on the grounds that any free variable used to express generality can simply be universally quantified, replacing the "any" by "all." As we have seen, that is not exactly correct, since it leaves no way of expressing certain facts that cut across types and orders within the formalism.[17]

§2. Cantor

The paradoxes posed no problem for Cantor's theory of sets—transfinite objects that can be counted. Indeed, in 1904 Cantor queried Jourdain about the availability of the second volume of Russell's *Principles.* Jourdain replied that it would not be available for some time, since Russell wished to present a "solution" of his "contradiction" in it, and he had not yet found one. Cantor replied with a discussion of the "difficulty" that Russell had described: Russell slightly extended Cantor's proof that $2^{\aleph_0} > \aleph_0$ to show that $2^{\mathfrak{a}} > \mathfrak{a}$ when $\mathfrak{a}$ is the cardinality of any *set* $\mathfrak{M}$. The extended proof shows, given a *set* $\mathfrak{M}$, how to form a *totality* $\mathfrak{G}$ of greater power. But Russell tried to apply the proof

[16] Hylton has argued [Hyl90, pp. 152–154] that Russell could not have employed schematic letters, because his conception of logic as universal blocked anything like an ascent to a metalanguage. Perhaps that is why Russell's pronouncements on "any" were confusing.

[17] Hylton has argued that the inability of the formalism of the *Principia* to express its own formulation is yet another fatal blow to Russell's logicism, given Russell's conception of logic as universal [Hyl90, pp. 159–161]. Even if the notion of "any" turns out to be compatible with Russell's conception of logic, I do not know whether it is strong enough to enable the formalism to be used to formulate itself.

with an inconsistent multiplicity or totality in place of $\mathfrak{M}$. But since this $\mathfrak{M}$ is not a set, a totality corresponding to $\mathfrak{G}$ cannot be formed, and no contradiction arises.[18] Thus, in Cantor's eyes, the difficulty is avoided. Russell had said [Rus03, p. 368] that "the application of Cantor's argument to the doubtful cases yields contradictions." Cantor had never accepted those cases.

Cantor's notion of a set is that of a collection "defined by the enumeration of its terms," as Russell said (see §1). I shall refer to that as the *combinatorial* notion of a collection.[19]

$\mathcal{P}(A)$

Cantor started investigating combinatorial collections of exceptional points in order to extend the results of Fourier analysis to as many functions as possible, building on the general definition of a function usually attributed to Dirichlet. The work was part of a program of freeing analysis of the restriction to functions given by analytic expressions—that is, to functions given by rules.

The values of a function are determined by the collection of points that form the graph of the function. The logical notion of a collection, that is, the notion of a collection determined by a rule, therefore goes hand in hand with that of a function determined by a rule or analytic expression: a function is determined by a rule, an analytic expression, if and only if its graph is given by a corresponding rule, and the graph is therefore a logical collection.

We see that Cantor's work that led to his set theory and to the notion of a combinatorial collection grew in a natural way out of what amounted to the attempt to free mathematics of the restriction to logical, rule-based collections. The whole point of the combinatorial notion is that combinatorial collections may exist whose members *cannot* be characterized by any rule. The Cantorian notion of a combinatorial collection is not merely different from the Peano–Russell notion of a logical collection—it arose in opposition to it.

I have just presented combinatorial collections as more general than logical collections. There are two main marks of the additional generality: First, on any fixed infinite domain there are more combinatorial collections than logical collections. That bald assertion presupposes some suitable clarification of what a permissible rule is for forming a logical collection. After all, given any

[18] The full correspondence I have just described is published in [GG71, pp. 118–119]. I have translated the relevant passage from Cantor's letter as Appendix A.

[19] The term is suggested by [Ber35b, pp. 259–260], compare [Mad90, pp. 102–103].

Thell Game that V i.e. set of [φ is set, ∀x φ(x) is set] is more general than the set theoretic notion of set.

combinatorial collection C, one could consider the instruction to collect the members of C to be a permissible rule. In that case, every combinatorial collection is trivially a logical one. But for any specification of allowable rules of which I am aware that does *not* presuppose combinatorial collections, the assertion holds. Second, combinatorial collections obviously obey the Axiom of Choice, while it is at best dubious whether logical collections do. I shall, however, postpone a discussion to §V.1.

!!

There is a different sense in which things go the other way: combinatorial collections are restricted, while logical collections are not. Since combinatorial collections are by a rule enumerated, some multiplicities may be too large to be gathered into a combinatorial collection. We have already seen Cantor's example—the multiplicity of all ordinal numbers. In contrast, the size of a multiplicity seems absolutely irrelevant to whether it forms a logical collection. Since there is a property characterizing the ordinal numbers—just that of being ordinal numbers—it seems that they do form a logical collection. That is part of why Russell's theory of logical collections led to paradox while Cantor's theory of combinatorial collections did not. Any restriction on logical collections motivated by the notion of a logical collection would have to be a restriction on allowable rules, a restriction like that imposed by the ramified theory of types, not a simple restriction on size.

'set' but 'is set' theory'

Though Cantor's theory was free of contradictions, it had other problems. In order to make it clear what those problems were, I shall give an axiomatic reconstruction of what I take to have been Cantor's mathematical theory in the period from somewhere around the time he arrived at the ideas in the *Grundlagen* of 1883 to the time he realized that the ideas of [Can91], the paper in which he first published his "diagonal argument" (translated as Appendix B to this chapter), could be used to show that the continuum of real numbers had cardinality $2^{\aleph_0}$. That period encompassed the main part of the development of his theory. It began with his acceptance of the ordinal numbers as objects of study in their own right, and it ended at a time when he was to publish only two more works on set theory, albeit important ones. Those last two works show some awareness of the problems with the theory as developed earlier. Note that that period of development came before the diagonal argument that led to Russell's paradox.

As we have seen, Cantor did not work axiomatically. He was working out the facts on the basis of a picture or conception, not on the basis of stipulated assumptions. Nonetheless, he did take each of the principles I take as Cantorian axioms to be, in one or another sense, basic.

* 'Combinatorial collections' are a special case of allowable sets

I shall engage in some simplification. For example, Cantor did not take 0 to be an ordinal number; he started with 1. I shall start with 0 anyhow. Cantor identified things other than sets (and sometimes, it seems, perhaps sets as well) with their singletons.[20] It is not clear whether Cantor took either ordinal numbers or cardinal numbers to themselves be sets. I shall remain neutral about whether ordinal numbers are sets in the way that I reconstruct Cantor's view. I shall leave cardinal numbers out of account altogether. They do not play a central role in Cantor's view, and so adding them on is not especially illuminating. Besides, later developments have made it clear that initial ordinal numbers will serve perfectly well as cardinal numbers.[21] I shall give a definition of the least infinite ordinal number, ω, that does not rely on a prior knowledge of the natural numbers. Cantor tended to rely on such knowledge. I have also modified Cantor's notation for number classes.

Technical Remark. Cantor worked in German, not in a formal language, and he worked years before anyone had distinguished between first- and second-order logic. (See §V.3.) He made free use of the notions of function and relation, taking them to be part of an antecedently given background.

Russell (and Frege before him) could not have adopted Cantor's procedure: Russell's notion of a class is so intimately connected to that of a relation that to introduce it without carefully specifying what is meant by a relation would have begged the question, and an analogous comment applies to Frege. But Cantor's notion of a set is a quite different one, as I have emphasized beginning in §III.4, and his notion is sufficiently far from those of relation and function that the present procedure is a reasonable one. It is true that one of the principal reasons that Cantor introduced set theory was to understand real-valued functions. But he did not regard the general notion of a function as problematic—his problems concerned special aspects of real-valued functions. Cantor's background theory of functions could not have led him into paradoxes in the way

[20] A singleton is a set with one member. Given an object a we can form the singleton $\{a\}$, a set that has a as its only member. Since a need not be a set, it is clear that a and its singleton need not be the same, and we shall, in general, take them to be distinct.

[21] An ordinal number α is an *initial ordinal number* if α is no greater than any ordinal number β such that the set of predecessors of β has the same power as the set of predecessors of α.

that Russell's and Frege's theories led them into paradoxes, since Cantor only considered functions on specified domains, and he never considered domains that consisted of functions of any very general sort. The problems of circular reference and self-application were therefore far from anything he considered.

When I claim that something or other follows from Cantor's axioms, it becomes necessary to specify what the assumptions of the background second-order logic are. All that is assumed is the usual quantifier rules plus the axiom scheme of comprehension for first-order formulas with parameters—which says that such formulas define legitimate relations and functions, not sets! (See, for example, [Sha91, p. 66].)

Here are the Cantorian axioms:[22]

Axiom 2.1. *The ordinal numbers are linearly ordered by* $<$.

Axiom 2.2. *There is a least ordinal number, 0.*[23]

Axiom 2.3. *Every ordinal number* α *has an immediate successor* $\alpha + 1$.[24]

Axiom 2.4. *There is an ordinal number* ω *such that* $0 < \omega$; *for every ordinal number* α, *if* $\alpha < \omega$, *then* $\alpha + 1 < \omega$; *and for every nonzero ordinal number* $\alpha < \omega$ *there is an ordinal number* β *such that* $\alpha = \beta + 1$.

[22] The less mathematically sophisticated reader may wish to skip directly to the discussion of the axioms. Axioms, definitions, theorems, and lemmas will be numbered in a single numbering system within each section of the book. Thus, the number 2.1 in Axiom 2.1 is the first numbered item of § 2 of the present chapter. A reference to, for example, Axiom 1, is always a reference to the first numbered item in the present section; a reference to Axiom 2.1 will be a reference (made outside of § 2) to the first item of § 2 of the present chapter; and a reference to Axiom IV.2.1 (made outside of Chapter IV) is a reference to the first numbered item of § 2 of Chapter IV. The system has the virtue that it is easy to locate items to which reference has been made. It has the defect that the numbering does not always reflect the logical grouping of the numbered items.

[23] More precisely, every ordinal number α other than 0 is such that $0 < \alpha$.

[24] More precisely, the axiom says that for every ordinal number α, there is an ordinal number β such that $\alpha < \beta$ and for any ordinal number γ, if $\alpha < \gamma$, then $\beta \leq \gamma$. It is easily seen that the ordinal number β is unique. Call it $\alpha + 1$.

That is, ω is a nonzero ordinal number whose predecessors are closed under successor and whose nonzero predecessors all have predecessors.

Definition 2.5. A *set* is the range of a one-to-one function with domain a proper initial segment of the ordinal numbers.

The definition says that a set is whatever can be counted. Note that it follows from the definition that the predecessors of any ordinal number form a set.

Axiom 2.6 (Extensionality). *Sets with the same members are equal.*

Axiom 2.7. *Every set of ordinals has a least upper bound.*

Note that it follows from the axiom (and Definition 5) that every proper initial segment of the ordinal numbers is the set of predecessors of an ordinal. Using Definition 5 and Axiom 7 we see that ω is the least nonzero ordinal number with predecessors that are closed under successor.

Axiom 2.8. *For every ordinal number* α *there is an associated set* (α), *the* number class *of* α, *such that* β *is in* (α) *if and only if* β *is an ordinal number and the set of predecessors of* α *is the range of a one-to-one function with domain the predecessors of* β.

A set can be counted by α if and only if it can be counted by any member of (α) and only by members of (α).

The above axioms serve to emphasize the primacy of ordinals in Cantor's conception of set theory: the only set-existence principles are Axiom 8, which postulates a set of ordinals, and the definition of set, which ties each set to an ordinal that counts it. The two set-existence principles cohere well: Since (α) is a set of ordinals, it has a least upper bound, say β, by Axiom 7. But when α is infinite, it is not hard to prove that (α) is the range of a function with domain the predecessors of β, and hence that (α) is a set in the sense of the definition. In fact, one can clean things up a bit by replacing Axiom 8 by

Axiom 2.9. *For every ordinal number* α, *there is an ordinal number* $\beta > \alpha$ *such that the set of predecessors of* β *is not the range of a one-to-one function with domain* α.

The resulting set of axioms is equivalent to the one above, but it has only one set-existence principle, Definition 5.

Either set of axioms requires supplementation by another principle that Cantor often in effect made use of, though he nowhere made anything like it explicit. Say that a one-to-one function F with domain the set of predecessors of an ordinal and with range S *witnesses* that S is a set. According to Definition 5, every set has a function witnessing that it is a set.

Axiom 2.10. *Let S be a set of sets, and let F be a function with domain the predecessors of some ordinal number α witnessing that S is a set, that is, such that every member of S is $F(\gamma)$ for some $\gamma < \alpha$. Then there is a binary function H such that for every $\gamma < \alpha$, the unary function $H(\gamma, \bullet)$ that remains after γ is plugged into H has domain the set of predecessors of an initial ordinal and witnesses that $F(\gamma)$ is a set.*

Axiom 10 is based on the rather natural principle that since every member of the set S has a function witnessing that it is a set, there is a set of such witnesses. Perhaps Cantor did not make anything like it explicit because it is a principle concerning functions, not sets, and he took himself to be working with an antecedently given notion of function.

Technical Remark. The mathematically sophisticated reader may be a bit bemused at this point, since Definition 5 and Axiom 10 are both closely related to the Axiom of Choice. But each seems to be independent of the other in the present setting, since the Axioms of Union and Power Set are absent. On the basis of what I have taken to be Cantor's axioms, Axiom 10 entails the Axiom of Union, as is not hard to check.

If we supplement my Cantorian axioms with the assumption that the ordinals are sets, for example by using the von Neumann ordinals, and assume (for simplicity) that there are no urelements, then the resulting second-order axioms are equivalent to second-order ZF minus the Foundation and Power Set Axioms plus the axiom that for every cardinal (initial ordinal) κ, a next cardinal κ^+ exists and the following axiom: for every set S there is a set T such that for every member s of S there is exactly one member of T that is a well-ordering of s with order type an initial ordinal. The final axiom entails the Axiom of Choice but does not follow from it on the basis of the first-order version of the theory

just described, that is, in the absence of the Power Set Axiom.[25] Andrzej Zarach has investigated theories closely related to the one just outlined [Zar82].

The axiomatization given here relies heavily on the notion of a witnessing function, which does not appear in Cantor's work. It is my way of expressing that a set can be "counted," in Cantor's own terms. Of course, it is immediate from Definition 5 that every set can be well-ordered, and Cantor relied on the notion of well-ordering instead of on the witnessing functions used here. But that involved him in the following detour: He had both well-order types [*Anzahlen*] and ordinal numbers [*Zahlen*]. Each well-ordered set is associated with the ordinal number such that the predecessors of that ordinal number in the natural order (<) have the same well-order type as the well-ordered set. (See [Can83, p. 168] or [Can76, p. 72]. The English translation does not always distinguish between *Anzahl* and *Zahl*.) There was some point to that added complexity, since the existence of well-ordered sets of one or another order type was part of Cantor's argument for the reality of the corresponding ordinal numbers. For example, the ordinal number ω is introduced in terms of the order type of the sequence of natural numbers. I have simplified Cantor's theory by making the association between ordinal numbers and well-ordered sets directly via witnessing functions, avoiding the mathematically superfluous detour through well-order types.

Let me briefly document that each of the proposed axioms is in fact a principle that Cantor accepted. Each pair of page numbers here is a reference to the *Grundlagen*, the first to [Can83] and the second to the English translation [Can76].[26]

Axiom 1. The numbers are in a "natural succession" [168, 72], and they are "comparable to each other" [177, 77].

Axiom 2. Actually, as mentioned above, Cantor started with 1. He spoke of the ordinal numbers as "an extension or rather a continuation of the sequence of real[27] whole numbers [1, 2, 3, . . . , ν, . . .] beyond the infinite" [165, 70].

[25] Zbigniew Szczepaniak has shown that the Well-Ordering Principle does not follow from ZFC minus Power Set (ZFC$^-$) if ZFC$^-$ is consistent [Zar82, p. 339]. His proof establishes the claim in the text.

[26] All but the briefest of the translations are, except as noted, my own.

[27] Cantor distinguished between "*reellen*" numbers—the continuum—and "*realen*" numbers—genuine ones—which include not only the real and complex numbers but also

Axiom 3. Cantor's "*first principle of generation*" is "the principle of the addition of a unit to an existing, already formed number" [195, 87, translation from [Can76]].

Axiom 4. Cantor held that there is

> nothing objectionable in conceiving of a *new* number—we shall call it ω—which is to be the expression for this: that the whole domain [*Inbegriff*] (I) [of the positive real whole numbers $1, 2, 3, \ldots, \nu, \ldots$] be given in its natural succession according to law. (Similar to the way in which ν is an expression for this: that a certain finite type [*Anzahl*] of units is unified into a whole.) [195, 87]

Definition 5. First, I shall discuss the reason for only allowing *proper* initial segments of the ordinal numbers. In investigating the suprafinite numbers, Cantor said that "we will get ever farther ahead, never reaching an unsurmountable limit, but also attaining not even an approximate grasp of the absolute. The absolute can only be acknowledged, but never known," and that "the absolutely infinite sequence of numbers . . . seems to me in a certain sense a suitable symbol of the absolute" [205, 94, the first translation is essentially that of [Can76]].

Now on to the central point of the definition.

> The concept of a *well-ordered set* shows itself to be fundamental for the whole theory of sets. That it is always possible to bring any *well-defined* set into the *form* of a *well-ordered* set seems to me to be basic and rich in consequences and through its general validity an especially remarkable law of thought . . . [169, 72]

Though that quote provides convincing evidence, I must discuss a certain passage in the *Grundlagen* that is often taken to be evidence that Cantor's set theory is so-called *naive* set theory,[28] a contradictory theory that is chiefly distinguished by the fact that it has as a postulate a Comprehension Principle much like that of Russell. The passage is in an endnote for Section 1 of

the *Grundlagen*, not in the text, so it would be curious if it had the central importance often claimed.[29] The passage opens as follows:

> By "set" [*"Mannigfaltigkeit" oder "Menge"*] I understand in general every many that can be thought of as one, i.e., any domain of definite elements which by means of a law can be bound up into a whole . . . [204, 93]

If one takes a "law" to be something like Peano's "condition," Frege's "concept," or Russell's "propositional function," then the passage is a classic statement of a Comprehension Principle, and that is how it is usually taken. That reading is, perhaps, encouraged by the fact that the passage continues[30] "and I believe that in this I am defining something which is related to the Platonic εἶδος or ἰδέα . . . " But Cantor's typical use of the word *law* in the *Grundlagen* is "natural succession according to law," which suggests quite a different picture: a "law" is, for Cantor, a well-ordering or "counting," and so the passage suggests Definition 5, not a Comprehension Principle. That reading is strongly supported by the fact that the passage continues

> as well as to that which Plato in his dialogue "Philebus or the Highest Good" calls μικτόν. He counterposes this to the ἄπειρον, i.e., the unlimited, indeterminate, which I call the non-genuine-infinite, as well as to the πέρας, i.e., the limit, and explains it as an ordered "mixture" of the two latter.

Let me just string together some relevant quotations from the *Philebus*.[31] The "μικτόν" is the subject of the third quote. What is at issue here is Cantor's understanding, not Plato's. It is not clear what Plato meant or what Cantor made of it, but there is no question that numbering, not conditions, concepts, or propositional functions, is the central idea.

the transfinite ordinals. I have not preserved Cantor's distinction in the translations. The reader will have no trouble sorting things out.

[28] As I learned from [Moo82, p. 260], the term is due to John von Neumann, who, following Ernst Zermelo, attributed the theory to Cantor. See, for example, [Zer08b, p. 200] and [vN25, p. 394].

[29] A similar passage *is* the opening of Cantor's *Beiträge* of 1895 [Can95, p. 481], but, as I shall argue below, Cantor had modified his theory by that time. The passage there is ambiguous in much the same way the one described here is. For a discussion of that passage and an explicit statement that Russell, not Cantor, is the originator of naive set theory, see [Hal84, p. 38].

[30] I am here and below using the translation in [Can76].

[31] It is interesting to note that the *Philebus* was likely written in response to Eudoxus, both his ethics and, what is relevant to the passages here cited, his theory of incommensurable ratios [Gos75, pp. 166–181, 196–206]. The translations are taken from [Gos75].

> But one should not attribute the character of indeterminate to the plurality until one can see the complete number between the indeterminate and the one. Then one can consign every one of them to the indeterminate with a clear conscience. [16D]
>
> If a person grasps any one, then, as I say, he must not turn immediately to its indeterminate character but rather look for some number. Similarly the other way round, when one is forced to start with what is indeterminate, one should not immediately look to the unitary aspect, but again note some number embracing every plurality, and from all these end up at the one. [18A]
>
> That of equal and double, and whatever puts an end to opposites being at odds with each other, and by the introduction of number makes them commensurate and harmonious. [25E]

Axiom 6. The Axiom of Extensionality, which is often taken to be constitutive of the notion of set, is curiously difficult to locate in Cantor's writings. Extensionality is present in the beginning of the passage discussed immediately above, in the idea that a set is a "domain of definite elements." The clearest statement is perhaps, as Hallett suggests, to be found in an article published in 1887 [Can87, p. 387] in which Cantor writes of a set "consisting of clearly differentiated, conceptually separated elements $m, m', \ldots$ and which is thereby determined and delimited" [Hal84, p. 34].

Axiom 7. Cantor defined [196, 87] the "*second principle of generation* of whole real numbers" to be that "if any definite succession of defined whole real numbers is given, of which no greatest exists, then on the basis of this second principle of generation a new number is created, which is thought of as a *limit* of those numbers, i.e., which is defined as the number next greater than all of them." As Cantor's application of the second principle of generation in the *Grundlagen* makes clear, a "succession" is intended to be an initial segment of the ordinal numbers. It is not clear what "definite" means, but the applications make it fairly clear that the principle is intended to apply to initial segments that are sets.

Suppose that given any set of ordinals one can form the set of all ordinals less than or equal to any member of the set. The set formed will be an initial segment, and so the second principle of generation guarantees that it has a least upper bound, which is therefore a least upper bound of the original set. The axiom then follows from the second principle of generation. The axiom is frequently more convenient to apply than the second principle of generation since it does not require forming an initial segment as an intermediate step.

The argument I just gave for Axiom 7 rested on assuming that given a set of ordinals one can form the set of all ordinals less than or equal to any member of the set, or, what is the same thing, that one can form the union of the members of the set of sets of predecessors of members of the first set. Cantor expressed no such idea in the period under discussion, but it—or something like it—seems to be required in order to make sense of Cantor's assertions about number classes and powers. Cantor rarely mentioned any number class beyond the first three explicitly. Though he explicitly declared that for every ordinal number γ there is a γth power [205, 94], he did not even make explicit mention of the ωth power in print until 1895 [Can95, p. 495], when he promised to prove its existence. He never did so in print. In 1899, in the letter to Dedekind in which he proved that the ordinal numbers form an inconsistent multiplicity [Can32a], he introduced $\aleph_\omega$, the ωth power, as the cardinality of the set of predecessors of the least ordinal number that does not have $\aleph_\nu$ predecessors for any finite ν and, what is essentially the same, also as the cardinality of the union of the first, second, third, and so forth, number classes. Thus, he allowed forming the union of a set of sets of ordinals when the result is an initial segment of the ordinals.[32] Such a union is just what is needed to justify my assumption. A more cautious reconstruction might replace Axiom 7 with "Every initial segment of the ordinals that is a set has a least upper bound." But, by a lemma of Azriel Levy [Lev68, p. 763], the resulting system has the full Axiom 7 as a consequence. I have used the apparently stronger version largely for perspicuity, but also because, as I just argued, Cantor seemed to use something like it, and he almost certainly did not know Levy's derivation of it.

Axiom 8. Along with the two principles of generation, Cantor stated a third principle, which he called a "stopping or confining principle" [*Hemmungs- oder Beschränkungsprinzip*]. He said that the principle or condition satisfied by all the ordinal numbers defined to that point in the *Grundlagen* is that their predecessors can be placed in one-to-one correspondence with the natural numbers, that is, that they are of the first infinite power. He went on to

[32] He used a denumerable disjoint union, while my proposed analysis allows an arbitrary union. He characteristically preferred disjoint unions, but he did consider nondisjoint unions from time to time [Can82, p. 152; Can84, p. 226], and so I think there can be no objection to allowing that in my reconstruction. Similarly, though he usually employed only finite or denumerable unions, he did make use of unions that cannot always be finite or denumerable to define multiplication of ordinal numbers [170, 73] and of cardinal numbers [Can87, p. 414].

define the second number class, which I have called (ω), as "*the domain of all numbers α that can be formed with the help of the two principles of generation*" such that the set of predecessors of α is of the first infinite power. (See [197, 88].) Now all that only gives the second number class, but he said earlier [167, 71], "In the same fashion the third number-class yields the definition of the third power, or the power of the third class, and so on."[33] Moreover, he referred to the third number class at a couple of other points in the paper.

Axiom 10. Axiom 10 is the hardest to defend, since Cantor nowhere stated anything like it. Various alternative formulations may be just as appropriate as the one I have given. But something like the proposed axiom is needed to prove the theorems Cantor does. In modern set theory, the counterparts of Definition 5 and Axiom 10 are equivalent—they are variants of the Axiom of Choice. But even when Cantor came to have doubts about Definition 5 (see below), he continued to use something like Axiom 10 unhesitatingly, which provides evidence that there was a principle at work that Cantor thought of as independent of Definition 5.

The first published theorem that cannot be rigorously proved without some form of the Axiom of Choice seems to have been a theorem of analysis proved by Cantor, published by his colleague Heine in 1872 [Moo82, p. 14].[34] There was no apparent recognition that a new principle was involved in the proof. Many theorems of set theory that Cantor subsequently published require some form or other of the Axiom of Choice. Virtually all are stated without proof, usually as elementary lemmas involved in the proof of other theorems. (See [Moo82] for a thorough survey.) Many of those theorems are immediate consequences of Definition 5. The theorem that every infinite set S has a denumerable subset provides an example: Let F witness that S is a set, and say the domain of F is the predecessors of α. Then $\alpha \geq \omega$, since S is infinite. Let F' be F with domain restricted to the predecessors of ω. Then the range of F' is a denumerable subset of S. Cantor did not publish that theorem until 1895. But he did state earlier that the powers are well ordered, that a set is finite if it has no proper subset equal in power to itself, and other immediate consequences of Definition 5.

In addition to the results just outlined that flow from Definition 5, Cantor also made use of the following results in the years indicated, which he regarded as too elementary to require explicit proof.[35] I have written them in modern notation, using $\sim$ for equivalence in power, $\cap$ for intersection, and $\bigcup$ for indexed union. (Thus, for example, $\bigcup_{i<\omega} A_i$ is the set of a such that for some $i < \omega$ a is in A_i.)

[33] The translation is taken from [Can76].

[34] Here is the theorem: A function f from the real numbers to the real numbers is continuous at a point p if and only if it is sequentially continuous at p.

Theorem 2.11 (1878, 1880). *Suppose that $A_i \sim B_i$ and $A_i \cap A_j = B_i \cap B_j = \varnothing$ for all $i < j < \omega$. Then $\bigcup_{i<\omega} A_i \sim \bigcup_{i<\omega} B_i$.*

Theorem 2.12 (1878, 1882, 1883, 1884, 1885). *A finite or denumerable union of finite or denumerable sets is finite or denumerable.*

Theorem 2.13 (1885). *A finite or denumerable union of sets of cardinality $\aleph_\alpha$ has cardinality $\aleph_\alpha$.*

Theorem 2.14 (1887). *Suppose that for all $i \neq j \in S$, $A_i \sim B_i$ and $A_i \cap A_j = B_i \cap B_j = \varnothing$. Then $\bigcup_{i\in S} A_i \sim \bigcup_{i\in S} B_i$.*

My argument for attributing to Cantor a commitment to Axiom 10 is simply that those results do not, so far as I can see, follow without it and that they have natural direct proofs with it.

Technical Remark. I shall prove Theorem 12 as an example: Let S be a finite or denumerable set of finite or denumerable sets. If any of the sets involved is finite, without loss of generality add more members to make it denumerable. By Axiom 10 there is a function H of two variables, each of which ranges over the natural numbers, such that the range of H is the union of the members of S. Let f and g be functions such that every pair of natural numbers (m, n) is of the form $(f(x), g(x))$ for exactly one natural number x. (In 1874 Cantor in effect showed that such functions exist, in the paper in which he proved that the algebraic numbers are denumerable.) The function $H(f(x), g(x))$ witnesses that the union of the members of S is a finite or denumerable set, as required, if the members of S were pairwise disjoint. Otherwise, one must delete duplications from the range of the function to obtain the required witness. The proof simultaneously shows that the union is a set and that it is finite or denumerable.

[35] See [Moo82, pp. 30–37] for references and a detailed discussion.

I have now introduced my reconstruction of Cantor's theory and argued for it. It is therefore time to admit that it does not suffice for formalizing Cantor's work during the period indicated. Such a formalization requires the additional assumption that the real numbers form a set and also, perhaps, the additional assumption that the functions from the real numbers to the real numbers form a set. (I say perhaps because the latter assumption is used only for offhand remarks that occur in endnotes and the like.) Cantor can be argued to have derived the additional assumptions from a guiding principle that is not a part of my reconstruction, perhaps one that says something like "every domain of a mathematical variable is a set." I shall call that the Domain Principle.[36] Indeed, some such assumption forms part of Cantor's argument for the existence of ω, as is indicated by the quote used to justify Axiom 4 above. The Domain Principle is not suitable for axiomatic formulation, but that is not why I have omitted it from account.

During the period under discussion, Cantor believed that he would be able to prove on the basis of his theory, essentially as I have reconstructed it above, that every mathematical domain is a set. Thus, the Domain Principle, or anything like it, was superfluous as an additional mathematical assumption. In particular, Cantor believed that he would be able to prove that the real numbers and the functions from the real numbers to the real numbers form sets within a framework like the one I have outlined, even though he could not yet even prove that the real numbers were not an absolutely infinite totality. Let us see, following [Hal84, pp. 74–81], why he thought so.

Cantor first showed that there are more real numbers than natural numbers or algebraic numbers in 1874. In 1891 he proved that $\aleph_0 < 2^{\aleph_0}$. (See Appendix B.) Thus, when he noted in 1895 that the real numbers have cardinality $2^{\aleph_0}$ [Can95, p. 488], the result of 1891 yielded a new proof that there are more real numbers than natural numbers. The later proof is the one in common use today. In fact, the earlier proof is no longer well known to mathematicians.[37] It was the earlier proof that shaped Cantor's thinking in the period under consideration.

[36] Hallett formulated what he calls the "domain principle" or "principle (a)" in much the same spirit, though his principle is not identical to mine. He argued along more or less the indicated lines in some detail.

[37] The basic technique of the earlier proof (1884) is still the one used to prove Cantor's theorem that no denumerable set is perfect [Can84, p. 215]. Since the real numbers form a perfect set, it follows that the real numbers are nondenumerable.

Here is the earlier proof: It suffices to show that given any sequence $u_1, u_2, \ldots$ of real numbers there is a number that is not in the sequence in any interval of real numbers (a, b)—that amounts to showing that any sequence $u_1, u_2, \ldots$ that supposedly lists all of the real numbers, thereby showing them denumerable, must omit at least one in every interval. Let a' and b' be the first two members of the sequence $u_1, u_2, \ldots$ that are in the interval (a, b), with $a' < b'$; let a'' and b'' be the first two members of the sequence that are in (a', b'), with $a'' < b''$; and so forth. If the sequence of nested intervals only goes on for finitely many steps, then there is at most one member of the last interval that appears in the sequence $u_1, u_2, \ldots$. Any other member of that interval is omitted from the sequence, as required. If the sequence of nested intervals is infinite, then the left endpoints converge to a point a^∞ and the right endpoints to b^∞. If $a^\infty = b^\infty$, then that point is as required. Moreover, as Cantor noted, if $u_1, u_2, \ldots$ is an enumeration of all the algebraic numbers, then that will be the case, and so a^∞ will not be algebraic. Finally, if $a^\infty < b^\infty$ then any member of $[a^\infty, b^\infty]$ (including the endpoints) is as required. That completes the proof.

Next, I shall present Cantor's proof in the *Grundlagen* that there are more numbers in the second number class, (ω), than there are natural numbers, or, as Cantor actually put it, that the second number class is of greater power than the first number class. Note how similar this proof is to the proof that there are more real numbers than natural numbers. (I have changed Cantor's notation to emphasize the parallelism between the two proofs.)

It suffices to show that given any sequence $u_1, u_2, \ldots$ of members of (ω) there is a member of (ω) that is not in the sequence. Let a be the first member of the sequence, let a' be the first member of the sequence that is greater than a, let a'' be the first member of the sequence that is greater than a', and so forth. If the increasing sequence $a, a', a'', \ldots$ only goes on for finitely many steps, then its last member is the greatest number that occurs in the sequence $u_1, u_2, \ldots$. Its successor is as required. If the increasing sequence $a, a', a'', \ldots$ is infinite, there is a least ordinal number greater then all its members, call it a^∞, which is in (ω) (since its predecessors have in effect been specified as a denumerable union of finite or denumerable sets). Thus, the ordinal number a^∞ is as required. That completes the proof. (Actually, Cantor went on to show that the ordinal number I have called a^∞ is the least upper bound of $u_1, u_2, \ldots$.)

Each of the two proofs shows that a set does not have the power of the natural numbers by showing how, given a sequence of members of the set, to find a member of the set that is not in the sequence. Moreover, that is done in each proof by considering appropriate subsequences, and in the main cases the number that is outside the sequence is obtained as a "limit" of the subsequence. Given the strong parallels between Cantor's analysis of the real numbers and his analysis of the second number class, it is small wonder that Cantor believed that the real numbers and the second number class (ω) were intimately connected and that he would eventually be able to prove a strong form of the Continuum Hypothesis: the real numbers have the power of the second number class. Note that a proof that the real numbers have the power of the second number class would show that they form a set, according to Definition 5, and so obviate the need to apply a domain principle. That was, I believe, the chief importance of the Continuum Hypothesis for Cantor—it would show that the real numbers form a set, and hence that they were encompassed by his theory.

Cantor repeatedly thought that he had proved that the real numbers have the power of the second number class, and he announced countless times that he hoped to publish such a proof.[38] He also announced that the functions from the real numbers to the real numbers have the power of the third number class [207, 95]. That would once more make an appeal to a Domain Principle unnecessary by showing that the functions from the real numbers to the real numbers form a set.

Cantor took the analogy between the real numbers and the second number class very seriously. In §9 of the *Grundlagen* he outlined three methods of introducing the real numbers: Weierstrass's method relying on series, Dedekind's method relying on cuts, and his own method relying on sequences. He gave various reasons to prefer his own over the others, including this one: his, unlike the others, generalizes to the case of transfinite numbers [190, 84]. What he had in mind is that just as irrationals are introduced to play the role of limits of Cauchy sequences of rational numbers, new ordinal numbers are introduced to play the role of limits of sequences of ordinals. That is strikingly clear in this passage:

> Indeed ω can in a certain way be viewed as the limit that the variable finite whole number ν aims at, though only in the sense that ω is the *smallest* transfinite ordinal number, i.e., the smallest *fixed* number that is greater than *all* finite numbers ν; in the same way $\sqrt{2}$ is the limit of certain variable, increasing, rational numbers, though here in addition the difference between $\sqrt{2}$ and these approaching fractions becomes arbitrarily small, whereas $\omega - \nu$ always equals ω; this difference does *not* alter that ω is to be recognized as just as definite and complete as $\sqrt{2}$, nor does it alter that ω has in it some traces of the numbers ν that aim at it, just as $\sqrt{2}$ does of the approaching rational fractions.
>
> The transfinite numbers are in a certain sense themselves *new irrationalities* and in fact in my opinion the best method of defining the *finite* irrational numbers is wholly similar to, and I might even say in principle the same as, my method described above of introducing transfinite numbers. One can say unconditionally: the transfinite numbers *stand or fall* with the finite irrational numbers; they are like each other in their innermost being; for the former like the latter are definite delimited forms or modifications (ἀφωρισμένα) of the actual infinite. ([Can87, pp. 395–396], my translation, but see also [Hal84, p. 80])

Cantor had produced a simple but powerful theory in which he could formulate a lot of new and interesting mathematics, but a proof that the real numbers have the power of the second number class—a proof that they form a set—continued to elude him. In 1891 he published a new proof that there are infinite powers other than that of the natural numbers, a proof that did not "depend on considering the irrational numbers."[39] What he had in fact shown is that for any set L and some fixed pair of distinct elements, the set of functions from L to that pair has power strictly greater than that of L. Thus, the infinite powers have no maximum, a result he had shown quite differently in the *Grundlagen*. First, he stated and proved the theorem in the case in which L is the set of natural numbers. Next, he stated the fully general form of the theorem that I just did, but he did not give a proof. Instead, he illustrated the method by showing that the theorem holds in the case in which L is the set of real numbers between 0 and 1. To avoid misunderstandings, let me state explicitly that there is no mention in the article of any set of subsets of a set, and that there is no proof in the article that there are more real numbers than natural numbers (though Cantor does mention that that was proved in an earlier article).

Cantor had reason for wishing to avoid any use of the irrational numbers in

[38] See, for example, [192, 86]. Detailed histories may be found in [Hal84], [Dau79], and [Moo82], among others.

[39] See Appendix B for a full translation of his article.

proving the theorem that there is an infinite power other than that of the natural numbers, but one should notice that he already had a proof that avoided the irrational numbers: the proof in the *Grundlagen*. I can only think that, since he thought of the transfinite numbers as "new irrationalities," he thought of that proof as involving irrational numbers in an extended sense. Thus, what is important is that the new proof avoids both the irrational and the transfinite numbers.

As Joseph Warren Dauben noted, the irrational numbers were controversial [Dau79, p. 165], and so, of course, were the transfinite numbers. Kronecker in particular was an influential figure who denied the existence of irrational numbers. (See, for example, [Dau79, p. 69].) It was therefore worthwhile, for polemical purposes, to avoid their use in proving that there is more than one infinite power. The two earlier proofs of that showed that sets defined in a similar way were large—sets defined by introducing limits to sequences or successions. But Cantor had no mathematical proof with which to confront his opponents that one could legitimately introduce such limits either in the case of the real numbers or in the case of the second number class. In the case of the irrational numbers he seems only to have had the argument that in analysis numbers present themselves in the form of limits of sequences.[40] In the case of the transfinite numbers he used the second principle of generation discussed above. Whatever the justification for those assumptions that suitable sequences have limits—and Cantor clearly believed them to be justified—a proof that there are infinitely many infinite sizes that is independent of them would clearly be more convincing than one that did depend on them. Moreover, an independent proof of that result could serve to bolster those assumptions.

The new proof is independent of any notion of limit or of transfinite number. It does, however, require a new set-existence principle: if L is a set, then so is the domain of all functions from L into an arbitrarily fixed pair. That principle can clearly be justified using the Domain Principle, but unlike the Domain Principle it is mathematically precise and provides a basis suitable for proving theorems.

As Russell had realized by 1900 (see [Cof79, p. 33] and [Rus03, p. 366]), any function from a set L to a pair is fully determined by the subset of L that is taken by the function to a fixed member of the pair: in effect, we think of the two members of the pair as meaning "yes, this element is in the subset" and "no, this element is not in the subset." The affinity with Russellian propositional functions is obvious. Thus, the domain of all functions from L into some fixed pair is canonically identifiable with the domain of subsets of L. It follows that the new principle is equivalent to the Power Set Axiom in common use today: the subsets of a set form a set. I shall therefore, at the cost of a slight anachronism, refer to the Cantorian principle above as the Power Set Axiom.

By 1895 [Can95, p. 488], Cantor had realized that the Power Set Axiom had another important consequence: he could show that the power of the set of real numbers (the continuum) was that of the set of functions from the natural numbers to a pair or, as he now wrote, that $\mathfrak{c} = 2^{\aleph_0}$. The proof of that fact is sufficiently easy, given what Cantor knew, that it would surprise me if he had not already discovered it in 1891, but the exact date doesn't matter.

The point is *not* that the new consequence that $\mathfrak{c} = 2^{\aleph_0}$ provided a new proof of the nondenumerability of the set of real numbers when combined with the 1891 result that, in the 1895 notation, reads $\aleph_0 < 2^{\aleph_0}$. Though that is the Cantorian proof familiar today, Cantor never actually gave it. Rather, the point is that the new Power Set Axiom enabled Cantor to prove for the first time that the real numbers form a set, instead of just taking that as an additional assumption. Cantor may well have seen that as a victory of consequence for more than just the extension and development of his theory: it provided at least the beginning of a new independent argument for the existence of the real and transfinite numbers. The Power Set Axiom had become vital for Cantor.

The Power Set Axiom, however, was not easily integrated with the conception of a set as anything that can be counted. For the first time, Cantor needed to allow for the existence of a set that he did not know how to introduce explicitly via a counting, or, more precisely, that he did not know how to well-order in a definable way. (I am excepting Cantor's earlier use of the real numbers—though he did not know how to count them, he had good reason to think that he would eventually be able to do so.)[41]

[40] Cantor didn't actually say that. What he did do is take it as an argument against Dedekind's definition of irrational real numbers in terms of "cuts" and as an argument in favor of his own definition in terms of sequences that numbers do *not* present themselves in the form of cuts [185, 81].

[41] As mentioned in §III.2, as a result of considering the real numbers Cantor did very briefly doubt that every set can be well ordered. That supports the point that Cantor's

Technical Remark. Since I am talking here as if Cantor were only interested in *definable* well-orderings, let me point out an ambiguity in Cantor's theory. The axioms are second order, and I have so far just been acting as if "function" meant arbitrary function in the sense of today's standard second-order logic, and so forth. But given the lack of clarity of Cantor and his contemporaries about definability, it would not be terribly implausible to reinterpret what Cantor had in mind as involving not arbitrary functions in the modern sense but instead α-recursive functions for some α. (See, for example, [Hin78, p. 377] for a definition and discussion.) In that case, Cantor's theory becomes the theory of α-finite sets for an infinite, recursively regular (that is, admissible), recursively inaccessible ordinal α.[42] This idea has its attractions for Cantor exegesis. For example, his "unconscious" use of a principle like Axiom 10—a choice principle—no longer involves an additional assumption, and his great interest in notations for ordinals ("normal forms" for denumerable ordinals and the like) becomes better motivated. Moreover, in the *Grundlagen* and elsewhere, Cantor gave a number of recursive definitions of larger and larger initial segments of the recursive ordinals and seemed to define the second number class as consisting of the set of those. That is, of course, easily explained away, since he could hardly have given nonrecursive examples, but it does cohere well with the present proposal.

Cantor always intended his theory of sets to be, in some none-too-clear sense, as comprehensive as possible. That strongly militates against the kind of interpretation of his work given in this Remark. His extra-mathematical intentions went beyond this interpretation and in that sense rule it out. The point is rather that up until 1891 nothing in his mathematical work even suggested that there might be any possibility of a set that did not have a definable well-ordering on it.

At some point after 1891 Cantor's mathematical thinking began to include the set of all functions from a set to two elements as an important example of a set, and so it was no longer part of his conception of sets that every set is born well-ordered. In Cantor's subsequent and final two publications on set theory, his *Beiträge* [Can95, Can97], the Power Set Axiom enters in in §4 [Can95, p. 487], when the exponentiation of cardinal numbers is defined, but many of the principles he had earlier taken to be obvious had become conjectures in need of proof [Moo82, pp. 44–46]. In particular, he no longer assumed that the powers are linearly ordered [Can95, p. 484]—that is "by no means self-evident and can hardly be proved at this stage." He announced in 1895 that he would show that the powers are even well-ordered [Can95, p. 495], presumably in the second (1897) article, but he did not. Since those results would easily follow from the principle that every set can be well-ordered, presumably he had come to doubt that too. Indeed, ordinal numbers and well-orderings had lost their primacy: in the *Grundlagen,* ordinal numbers were mentioned in the very first sentence, and well-ordered sets were defined in the second section, immediately after the introduction, but in the *Beiträge* well-ordered sets are not defined until §12, which is the beginning of the second article, and ordinal numbers not until §14. Cantor's theory was in trouble, but it was not trouble caused by the paradoxes. It was trouble caused by trying to fit the Power Set Axiom into a theory that took well-orderings to be primary.

In the letter to Dedekind that was discussed in §III.4, Cantor used his result that the ordinal numbers form an inconsistent multiplicity to show that the cardinal numbers are well-ordered, indeed to show that every cardinal number is an $\aleph$, that is, the power of the set of predecessors of an ordinal number: Suppose a multiplicity does not have any $\aleph$ as its cardinality. Then, Cantor wrote, the whole system of all ordinal numbers is projectible into that multiplicity. But then the multiplicity is inconsistent, and hence not a set.

Cantor must have had mixed feelings about the above proof. Jourdain discovered a similar one about which he wrote to Cantor. Cantor put his own in a letter to Jourdain, but subsequently refused Jourdain permission to publish the letter [GG71, pp. 115–118]. Of course, the proof would not have been necessary in Cantor's earlier theory. There one could just argue as follows: every set can be counted and hence has the same power as the set of predecessors of some ordinal number—indeed any ordinal number that can be used to count it. But the power of the set of predecessors of an ordinal is an $\aleph$.

A variant of the proof *does* show something about Cantor's earlier conception of sets that was to be important later. Namely, it shows that if a multiplicity is not a set, then it is larger than every set: Begin counting the multiplicity off. If you succeed at some ordinal stage, then the multiplicity is a set. But if you do not succeed, then all the ordinals will have been used without exhaust-

doubts about well-ordering and related topics were induced by the need to allow power sets—which are not canonically well ordered.

[42] The only remaining vestige of full second-order logic in this interpretation of Cantor's theory is that one must interpret the notion of a proper initial segment of the ordinals in a standard way, to guarantee that < is a well-ordering.

ing the set, that is, the multiplicity has a part that is the size of all the ordinals. Thus, the multiplicity is larger than every set. That is the origin of the later "limitation-of-size hypothesis": a collection forms a set just if it is not too large. Note that the limitation-of-size hypothesis arose within a theory that is free of all known paradoxes. It did not arise as a solution to paradoxes.[43] Indeed, on the logical conception of collection, which is the source of the paradoxes, it is not at all clear how size could be relevant to the question whether a multiplicity forms a set—the elements are, after all, not gathered, they simply obey a rule. That suggests that on the logical conception, one would have to limit not the size of collections but the structure of the rules, just as Russell did.

Once Cantor had accepted the Power Set Axiom as a second set-existence principle, he no longer had a unitary conception of set. He could therefore no longer say which multiplicities were sets; the only way he had to show that some were not sets was by arriving at contradictions. In the absence of a positive account, that policy seemed *ad hoc*. It is a corollary of the result that the cardinality of every set is an $\aleph$ that every set can be well-ordered or counted. But that no longer serves as a criterion for what things are sets, because one can now show, for example, that the set of real numbers can be well-ordered only after the fact—by showing that it is a set. One cannot show that it is a set by showing that it can be well-ordered. The elegant theory of the *Grundlagen* was lost, and it was not clear what could replace it.

§3. Appendix A: Letter from Cantor to Jourdain, 9 July 1904[44]

Today I want to reply to only one point in your kind letter, that is the difficulty, which Mr. Russell describes in his work *The Principles of Mathematics* pp. 365–368.

He starts out from my proof of the theorem

[43] The term *limitation of size* was, it must be admitted, introduced by Russell in 1906 to name a theory considered by him as one possible way of solving the paradoxes concerning logical collections. As he said [Rus05, p. 152], however, "This theory naturally becomes particularised into the theory that a proper class [that is, an allowable collection] must always be capable of being arranged in a well-ordered series ordinally similar to a segment of the series of ordinals in order of magnitude." That is, as we have seen, Cantor's theory developed before the paradoxes were known, and before Cantor had proved the existence of absolutely infinite collections.

[44] I have based my translation on the text as published in [GG71, p. 119].

$$2^{\aleph_0} > \aleph_0$$

which can easily be extended to the case where $\aleph_0$ is replaced by some transfinite cardinal $\mathfrak{a}$.

One assumes here some *set* (that is, a *consistent* multiplicity) $\mathfrak{M}$ with the cardinal number $\mathfrak{a}$ and imagines the totality $\mathfrak{S}$ of all coverings of $\mathfrak{M}$ with two mutually exclusive symbols, perhaps with 0 and 1.

The *elements* of $\mathfrak{S}$ are therefore definite coverings of $\mathfrak{M}$, each therefore an individual set of the same cardinal number $\mathfrak{a}$.

Were we now, as Mr. Russell proposes, to replace $\mathfrak{M}$ by an *inconsistent* multiplicity (perhaps by the totality of *all* transfinite ordinal numbers, which you call $\mathfrak{W}$), then a totality corresponding to $\mathfrak{S}$ could *by no means be formed.* The impossibility rests upon this: an inconsistent multiplicity because it cannot be understood as a *whole,* thus as *a thing,* can*not* be used as an *element* of a multiplicity.

Only *complete things* can be taken as *elements* of a multiplicity, only *sets,* but not *inconsistent multiplicities,* in whose nature it lies, that they can never be conceived as *complete* and *actually existing.*

§4. Appendix B: On an Elementary Question of Set Theory[45]

In the article "Über eine Eigenschaft des Inbegriffs aller reellen algebraischen Zahlen" (*Journ. Math.* **77**, 258) [(1874), see [Can32b, pp. 115–118]] one finds, probably for the first time, a proof of the theorem that there are infinite sets that cannot be placed into one-to-one correspondence with the totality of all finite whole numbers $1, 2, 3, \ldots, \nu, \ldots$, or, as I like to put it, that do not have the power of the number sequence $1, 2, 3, \ldots, \nu, \ldots$. From what has been proved in §2 it follows without further argument that, for example, the totality of all real numbers within any interval $(\alpha \ldots \beta)$ are *not* representable in the form of a sequence

$$\omega_1, \omega_2, \ldots, \omega_\nu, \ldots.$$

It is possible, however, to produce a much simpler proof of that theorem that does not depend on considering the irrational numbers.

Take any two symbols m and w that are distinct from one another. Now we consider a domain [*Inbegriff*] M of elements

$$E = (x_1, x_2, \ldots, x_\nu, \ldots),$$

[45] My translation of [Can91] as based on the reprinting [Can32b, pp. 278–280].

which depend on infinitely many coordinates $x_1, x_2, \ldots, x_\nu, \ldots$, such that each of these coordinates is either m or w. M is the totality of all elements E.

Among the elements of M belong, for example, the following three:

$$E^{\mathrm{I}} = (m, m, m, m, \ldots),$$
$$E^{\mathrm{II}} = (w, w, w, w, \ldots),$$
$$E^{\mathrm{III}} = (m, w, m, w, \ldots).$$

I claim that such a set M is not of the power of the sequence $1, 2, 3, \ldots, \nu, \ldots$.

That is shown by the following theorem: "If $E_1, E_2, \ldots, E_\nu, \ldots$ is any simply infinite sequence of elements of the set M, then there is always an element E_0 of M that corresponds to no E_ν."

To prove this, let

$$\begin{aligned} E_1 &= (a_{1,1}, a_{1,2}, \ldots, a_{1,\nu}, \ldots), \\ E_2 &= (a_{2,1}, a_{2,2}, \ldots, a_{2,\nu}, \ldots). \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \\ E_\mu &= (a_{\mu,1}, a_{\mu,2}, \ldots, a_{\mu,\nu}, \ldots). \\ &\cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot\ \cdot \end{aligned}$$

Here the $a_{\mu,\nu}$ are m or w in a definite manner. Produce now a sequence

$$b_1, b_2, \ldots, b_\nu, \ldots,$$

so defined that b_ν is *different* from $a_{\nu,\nu}$ but is also either m or w.

Thus if $a_{\nu,\nu} = m$, then $b_\nu = w$, and if $a_{\nu,\nu} = w$, then $b_\nu = m$.

If we now consider the element

$$E_0 = (b_1, b_2, b_3, \ldots)$$

of M, we see at once that the equality

$$E_0 = E_\mu$$

can be satisfied for no whole-number value for μ. Otherwise, for the μ in question and for all whole-number values of ν

$$b_\nu = a_{\mu,\nu},$$

therefore in particular

$$b_\mu = a_{\mu,\mu}$$

would hold, which is ruled out by the definition of b_ν. It follows immediately from this theorem that the totality of all elements of M cannot be brought into the form of a sequence $E_1, E_2, \ldots, E_\nu, \ldots$; we would otherwise be faced with the contradiction that a thing E_0 both is and is not an element of M.

This proof seems remarkable not only because of its great simplicity, but especially also because the principle that is employed in it can easily be extended to the general theorem, that the powers of well-defined sets have no maximum or, what is the same, that for any given set L another M can be placed beside it that is of greater power than L.

For example let L be a linear continuum, perhaps the domain of all real numerical quantities z that are $\geqq 0$ and $\leqq 1$.

Let M be understood as the domain of all single-valued functions $f(x)$ that take on only the two values 0 or 1, while x runs through all real values that are $\geqq 0$ and $\leqq 1$.

That M does *not* have *smaller* power than L follows from this: subsets of M can be specified that have the same power as L, *e.g.*, the subset that consists of all functions of x that have the value 1 for a single value x_0 of x and the value 0 for all other values of x.

But then M does *not* have the *same* power as L either. For otherwise M could be put into one-to-one correspondence to the variable z [of L], and thus M could be thought of in the form of a single-valued function

$$\phi(x, z)$$

of the two variables x and z, in such a way that through every specification of z one would obtain an element $f(x) = \phi(x, z)$ of M and also conversely each element $f(x)$ of M could be generated from $\phi(x, z)$ through a single definite specification of z. This however leads to a contradiction. For if we understand by $g(x)$ that single-valued function of x which takes only values 0 or 1 and which for every value of x is different from $\phi(x, x)$, then on the one hand $g(x)$ is an element of M, and on the other it cannot be generated from $\phi(x, z)$ by any specification $z = z_0$, because $\phi(z_0, z_0)$ is different from $g(z_0)$.

Since the power of M is neither less than nor equal to that of L, it follows that it is larger than the power of L. (Cf. *Crelle's Journal* **84**, 242) [(1878) [Can32b, pp. 119–133]].

I have already shown by entirely other means in my *Grundlagen einer allgemeinen Mannigfaltigkeitslehre* (Leipzig, 1883; *Math. Ann.* **21**) [Can83] that the powers have no maximum. There it was even proved that the domain of all powers, when we imagine them ordered according to size, forms a "well-ordered set" so that in nature for each power there is a next greater, and that for each endlessly increasing set of powers there is a next greater one that follows.

The "powers" represent the only and necessary generalization of the finite "cardinal numbers"; they are nothing other than the actually infinitely large "cardinal numbers," and they possess the same reality and definiteness as the finite cardinal numbers; the only difference being that the law-like relations among them, their respective "number theory," is partly different in kind from that in the region of the finite.

The further exploration of this field is a job for the future.

V

The Axiomatization of Set Theory

§1. The Axiom of Choice

It was crucial for, indeed constitutive of, Cantor's early theory that the Well-Ordering Principle hold, that is, that every set can be well ordered. (The term "the Well-Ordering Principle" is taken from [Moo82].) That principle had, with the advent of the Power Set Axiom, become questionable, a conjecture that it was critical to prove in order to rescue Cantor's theory of powers. At the very least, it was necessary to prove that the power set of a well-orderable set is well-orderable. The following test case, which is also the central case, suggests itself: Prove that the set of subsets of the natural numbers can be well-ordered, or, equivalently, prove that the real numbers can be well-ordered. That problem, together with the apparently closely related continuum problem, was the first on Hilbert's tremendously influential list of problems presented to the Second International Congress of Mathematicians in 1900. Since the Cantorian theory had proved inadequate, there was the closely related problem of arriving at a workable notion of set with which to formulate set theory. The ordinal numbers could not, it seemed, play the defining role Cantor had assigned them.

In 1904, Julius König gave a purported proof that the continuum cannot be well ordered, a result that would have put Cantor's theory in doubt. Zermelo found an error in the proof: König had relied on a theorem of Felix Bernstein, and Bernstein's proof is incomplete in the relevant case. (See, for example, [Moo82, pp. 86–88] for an excellent account of that often-told tale.) Within a month, in conversations with Erhard Schmidt, Zermelo had solved the problem of well-ordering the continuum to his own satisfaction and that of present-day set theorists [Zer04]. Zermelo's contemporaries were not so sure.

Zermelo's work presumed familiarity with Cantorian set theory. He used cardinal numbers and their products, functions from sets to their members, well-ordered sets, and the Power Set Axiom without discussion. He introduced the "assumption," which he later called the Axiom of Choice,[1] that for every set of nonempty sets there is a function that takes each of the nonempty sets to one of its elements and used that assumption to prove the Well-Ordering Principle.

It is characteristic of combinatorial collections that they obey the Well-Ordering Principle and the Axiom of Choice. They obey Well-Ordering because combinatorial collections are gathered by enumerating their elements, and they obey Choice because combinatorial collections are gathered by picking their elements in an arbitrary way, not necessarily in virtue of a rule. That makes it possible to pick one member out of each set in a set of nonempty sets. Such picking does not give rise to a single rule that selects exactly one member from each set in the set of nonempty sets: it may be done in an arbitrary way. Thus, Choice is dubious for logical collections, which require a rule.

Given a set M Zermelo applied his assumption to the set of its nonempty subsets to obtain a function γ. When m is any nonempty subset of M, he called $\gamma(m)$ the *distinguished* member of m. He defined a *γ-set* to be a well-ordered subset S of M such that each member a of S is the distinguished member of the set of elements of S that do not come before a. He then defined a *γ-element* to be a member of any γ-set, and showed that (i) the set L_γ of γ-elements is a γ-set, and thus well-ordered, and (ii) that $L_\gamma = M$, and hence that M is well-ordered. In support of his assumption, he said only that it is a "logical principle" that is "applied without hesitation everywhere in mathematical deduction." As an example, he noted that it is used to prove that if a set is decomposed into parts then there are not more parts than members of the set.

Zermelo's article provoked a storm of criticisms. Bernstein and Arthur Schoenflies objected to the proof that $L_\gamma = M$. Poincaré objected to the definition of L_γ. Jourdain claimed to have proved the result earlier, in a simpler way. Peano, Borel, Lebesgue, and René Baire objected to the assumption. I have just listed the published criticisms that had come to Zermelo's attention by 1908 [Zer08a]. There were other published criticisms [Moo78, p. 320].

[1] See [Moo82] for a thorough and careful discussion of precursors of the Axiom of Choice, equivalents of the axiom, early theorems whose proofs require the axiom, reactions to the axiom, and practically everything else about it.

Zermelo responded to his critics in a pair of papers [Zer08a, Zer08b], written within sixteen days of each other, that were virtually a single paper [Moo78, p. 319]. One of the papers [Zer08a] contains a new proof that every set can be well ordered, a detailed argument for the Axiom of Choice, and replies to critics. The other contains the first axiomatization of a Cantorian set theory.

Bernstein and Schoenflies were concerned, in somewhat different ways, with the "set" W of all ordinals and with Burali-Forti's paradox. Each wished to accept W as a set, and so they restricted set-theoretic principles in various ways in an attempt to block the paradox. In particular, they denied that, given any well-ordered set and an object not in the set, there is always a well-ordered set consisting of the members of the original one plus the new object ordered so that the new object comes after all the others. They pretty much had to do that, since the paradox arises immediately once one continues W with another object. They therefore criticized Zermelo for assuming that one could always extend a well-ordered set by a new element.

Zermelo's main reply was three pronged, though he also criticized the details of the theories of Bernstein and Schoenflies. First, he argued that any attempt to save W is pointless, since Russell's paradox shows that the problems that give rise to the Burali-Forti paradox go deeper than the theory of well-ordered sets and require "a suitable restriction of the notion of set" [Zer08a, p. 192], not merely a modification of the theory of well-ordered sets. Second, he claimed that set theories that admit W as a set will inherit its "inconsistent character" and that they are therefore doomed to failure. As evidence he cited Gerhard Hessenberg, who had noted that while Bernstein had used W to show that there are sets that cannot be well-ordered, Jourdain had used it to show the opposite. And third, and most important, he proved his result in a system in which W does not appear. The main point of giving his new proof is that that is even clearer than it was for the old proof:

> Already in my 1904 proof, having such reservations in mind, I avoided not only all notions that were in any way dubious but also the use of ordinals in general; I clearly restricted myself to principles and devices that have not yet by themselves given rise to any antinomy . . . Now I succeeded in completing my new proof without even the device of rank-ordering, and I hope thereby to have definitively cut off every possibility of introducing W. [Zer08a, p. 192]

We have just seen Zermelo's main motive both for his new proof and for his axiomatization of set theory. It was *not* to secure set theory from paradoxes.

(See [Moo78] and [Moo82, pp. 155–160].) Paradoxes were a side issue. It was to secure his theorem from the criticism that the methods he had employed in its proof led to paradoxes.

The old proof used, in addition to sets, the separate notion of orders that are imposed on sets. Today we just take such an order, or indeed any binary relation on a set S, to be the set of ordered pairs of members of S between which the relation holds. (Indeed, we take an n-ary relation to be a set of n-tuples for any n, and in particular we take a unary relation to be a set of members of S.) Since we know today how to reduce ordered pairs to sets, that device reduces talk of orders to talk of sets and so avoids postulating two sorts of objects. That avoidance is a serious convenience when trying to analyze what principles are being used. The device of ordered pairs was not available to Zermelo. What he therefore did in the second proof was reduce well-orderings on sets to particular sets by using an *ad hoc* device in order to avoid the need to postulate two sorts of objects, sets and orders. That is what made it possible for Zermelo to derive the Well-Ordering Principle from principles that concern only sets.

Zermelo was not attempting to present a theory of what sets are. He emphasized that he had sought out "the principles required for establishing the foundations of" set theory as it was historically given and that he would not discuss their origin in his article [Zer08b, p. 200].[2] It is not in fact quite correct that his principles are adequate for set theory as it was historically given, as we shall see below. What *is* true is that they are the principles required to prove Zermelo's theorem. Every single one of the axioms, with the exception of the last one, the Axiom of Infinity, is used in Zermelo's new proof, and of course the theorem depends on the Axiom of Infinity for its interest.[3] Zermelo's hodgepodge of axioms extracted opportunistically from a proof forms the basis of the axioms of present-day set theory. The method of obtaining the axioms was appropriate to Zermelo's limited purpose, but it should hardly be

expected to yield an axiomatization of a coherently motivated theory based on some distinctive conception of what a set is.

Poincaré rejected the actual infinite. He viewed the mathematics that is apparently concerned with the actual infinite as actually concerning the finite linguistic definitions that putatively describe actually infinite objects. He therefore thought of definitions of such objects as giving them existence, rather than pointing them out or otherwise distinguishing them from other objects. The set L_γ, which, recall, is a γ-set, is defined as the set of γ-elements, that is, all members of γ-sets. But then, since L_γ is a γ-set, an object could be a γ-element, and hence a member of L_γ, in virtue of its membership in L_γ. Thus, Poincaré viewed the definition of L_γ as viciously circular because it is, in his terminology, impredicative, and therefore incoherent.

Zermelo's reply was simple, but in the end decisive. He noted that impredicative definitions are common in analysis. They are indispensable to the practice of ordinary mathematics. Moreover, they are unobjectionable on Zermelo's view, since they do not create the objects they define, but merely distinguish them from other objects. Thus, L_γ is indeed a γ-set, and so something can be shown to be a member of L_γ in virtue of its membership in L_γ without circularity. Showing that something is in L_γ from the fact that it is in L_γ is not very informative, unless we have somehow identified L_γ using a different description. But there is nothing contradictory or viciously circular about Zermelo's definition if it picks out L_γ from a collection of previously existing objects, instead of creating L_γ. To summarize, impredicative definitions are necessary for ordinary mathematics, and they are unproblematic *if* one adopts a realist attitude about the objects defined, realist in just the sense that the objects exist in advance of the definitions, that they are picked out by the definitions, not created by them. That imposes a substantial constraint on any acceptable philosophy of mathematics.

Jourdain claimed that his proof that every set has an $\aleph$ as its power, outlined in §IV.2, accomplished all that Zermelo's did, but more simply. Zermelo pointed out that Jourdain's principles, which allow W, did not permit him to show that the continuum is a set. Thus, Zermelo's theorem entails that the continuum is well-orderable, while Jourdain's did not.

Zermelo went on to discuss Jourdain's proof. That proof made use of an arbitrary succession of choices—as Zermelo put it [Zer08a, p. 193], "after an arbitrary finite or infinite number of elements take an arbitrary element of the remainder as the next one; and continue in this way until the entire set is ex-

[2] Zermelo took "Cantor's original definition of a set" to lead to paradoxes, apparently mistaking it for a statement of Comprehension in the manner discussed in §IV.2. He thus took Cantor's conception to be naive. But Zermelo ignored what he took to be Cantor's definition and took his principles from Cantorian proofs. Zermelo's theory is thus extremely Cantorian despite his lack of recognition of that fact.

[3] The Union Axiom is used only to derive the Axiom of Choice—every set of nonempty sets has a choice function—from the Multiplicative Axiom—see the text below.

hausted." Though Zermelo had emphasized earlier in his article (p. 186) that the Axiom of Choice involved simultaneous choices, here he accepted Jourdain's use of a temporal succession of choices. Perhaps he found simultaneous "choices" just as dubious as Jourdain's successive ones, since he replaced the axiom that "a simultaneous choice of distinguished elements is in principle always possible for an arbitrary set of sets" with an equivalent that is less "tainted with subjectivity and liable to misinterpretation" [Zer08a, p. 186], namely what is now known as the Multiplicative Axiom: for every set T of pairwise disjoint nonempty sets, there is a subset S of $\bigcup T$ that has exactly one member in common with each member of T. The Multiplicative Axiom is, as Zermelo showed, equivalent to the Axiom of Choice. It was also stated, apparently independently, by Russell, who gave it its name. (I shall not discuss Russell's independent discovery of the Axiom of Choice in any detail. He was working within the theory of types—based on the logical notion of collection—when he discovered that he needed a new assumption—the Multiplicative Axiom—to fill in gaps in certain proofs. He used it reluctantly since he thought it to be complicated and dubious—as indeed it is as a principle concerning logical collections. See [Moo82, pp. 121–132].)

Zermelo objected not to successive choices but to Jourdain's assumption that the entire set will be exhausted, that is, the assumption that the set of well-orderings of subsets of the set has a maximal element. He said that that requires proof. The assumption is essentially a special case of Hausdorff's Maximal Principle,[4] which is equivalent to the Axiom of Choice, though Zermelo does not seem to have recognized that fact. Zermelo was certainly right that Jourdain had used an unidentified additional assumption. With that assumption, Jourdain's proof is still vitiated by its use of W. There is, however, a proof of the Well-Ordering Principle from Hausdorff's Maximal Principle that is simpler than Zermelo's proof of the principle from the Multiplicative Axiom. In 1932, in his editorial comments to Cantor's letter to Dedekind discussed in §§III.4 and IV.2, in which Cantor gave the same proof as Jourdain, Zermelo criticized the use of successive choices [Can32a, p. 117]: "the intuition of time is applied here to a process that goes beyond all intuition, and a fictitious entity is posited of which it is assumed that it could make *successive* arbitrary choices." Zermelo went on to say that one would have to use simultaneous choices.

Peano, Borel, Lebesgue, and Baire all doubted the assumption, that is, the Axiom of Choice. Peano essentially rested content with noting that the axiom did not follow from the laws of logic. Borel admitted that a version of the axiom that was restricted in its application to denumerable sets of nonempty sets might be acceptable. Lebesgue rejected any form of the axiom. Baire rejected any form of the axiom, and he rejected the Power Set Axiom (for infinite sets) as well. (See [Moo82, pp. 92–96].) Borel explicitly accepted Zermelo's proof, but he viewed it as demonstrating (the difficult direction of) the equivalence between the Axiom of Choice and the Well-Ordering Principle. Zermelo had, in Borel's eyes, shown two problems equivalent without solving either one [Moo78, p. 312].

Zermelo began his reply by carefully noting that he could not *prove* the Axiom of Choice, and that in accusing him of failing to provide a proof for it, his critics endorsed his own view of the matter. But [Zer08a, p. 187] "every proof in turn presupposes unproved principles." He said that to "reject such a fundamental principle" one would have to show it false or contradictory, and none of his critics had attempted to do so.

Zermelo mentioned two sorts of support for the axiom, or indeed any mathematical principle. The first is intrinsic support—that the axiom is "intuitively evident"—the second is extrinsic support—that it is "necessary for science."[5] That is in apparent contrast to his immediately preceding claim that his opponents could not reject a principle unless it were shown to be false or contradictory. I take it the point is this: A principle cannot be definitively rejected, and it may therefore be used, unless it is shown to be false or contradictory. But a principle must be accepted, and it *must* therefore be used, if it is "intuitively evident and necessary for science." Zermelo argued [Zer08a, p. 187] that Peano had selected his fundamental principles by analyzing what ones mathematicians have used "and by pointing out that the principles are intuitively evident and necessary for science." He noted that he could marshal the same kind of arguments for his axiom. The fact that it did not happen to appear in Peano's list is not an argument against it. Zermelo also pointed out that while his proof can be carried out in a system that is free of all known paradoxes, Peano's system is subject to paradoxes and hence inconsistent. It

[4] Hausdorff's Maximal Principle says that every partially ordered set has a linearly ordered subset that is ⊆-maximal among such subsets. (It might seem more natural to invoke Zorn's Lemma here, but Jourdain would not have viewed well-orderings simply as sets.)

[5] The useful terms *intrinsic* and *extrinsic* are taken from Maddy. See [Mad90, p. 118].

is simultaneously too narrow, in omitting the axiom, and too wide, in permitting paradoxes.

To show that the axiom is intrinsically motivated, indeed "self-evident," Zermelo noted [Zer08a, p. 187] that it had been, in effect, used by many mathematicians with a great deal of success "even though it was never formulated in textbook style." If a principle, possibly in variant forms, is independently and unquestioningly applied by many mathematicians, surely that shows that the principle is self-evident. Self-evidence is a psychological or perhaps sociological phenomenon, but, as Zermelo said,

> No matter if this self-evidence is to a certain degree subjective—it is surely a necessary source of mathematical principles, even if it is not a tool of mathematical proofs, and Peano's assertion that it has nothing to do with mathematics fails to do justice to manifest facts. [Zer08a, p. 187]

The reader will surely object that the axiom was not self-evident to, for example, Peano, Borel, Lebesgue, and Baire, but of that, more below.

Another objection that may occur to the reader is that while the Axiom of Choice is indeed self-evident for finite sets of nonempty sets, the extension to infinite sets is not self-evident. The appearance of self-evidence could therefore turn out to be simply a case of unwarranted generalization.[6]

The Axiom of Choice certainly is evident in the finite case. But it is not new in that case: the Axiom of Choice restricted to finite sets of nonempty sets is a theorem of, for example, Peano's system, as Zermelo noted [Zer08a, p. 187]. Zermelo noted that fact immediately before arguing for the self-evidence of the axiom. It is precisely the version that applies to infinite sets of nonempty sets that was at issue, and it was the earlier uses of that version that provided the basis for Zermelo's argument. Thus, the axiom provides an example of a positive principle, distinctively about the infinite, that is self-evident.

Because the importance of claims of self-evidence is often dismissed on the basis of an overly simple account of their function, it is worthwhile to emphasize the sophisticated nature of Zermelo's claim that Choice is self-evident. He did not at any time claim that the truth of Choice was revealed to him through some mysterious faculty or that it could be revealed to others

in that way. As Kitcher has emphasized [Kit88, p. 297], that is just not how mathematics grows. Zermelo did not ask us to accept the axiom on the basis of some armchair introspection. Nor did he treat self-evidence as a sufficient condition for acceptance. Zermelo's evidence for the self-evidence of Choice was that many mathematicians had used principles or techniques equivalent to applications of Choice (1) with a great deal of success and (2) without any awareness that a new principle was being applied.

There is a tendency to view the lack of awareness that a new principle is being applied as an oversight on the part of the mathematicians applying it. Zermelo did not have that tendency, and I believe that it is one that should be resisted. Those who employed Choice-like principles were not working in a late-twentieth-century axiomatic setting. They saw themselves as discovering truths, mostly about the real numbers, on the basis of what they already knew about the real numbers. A proof technique or principle, whether new or old, was fully appropriate to their task just if it is a *correct principle*. A correct principle is, more or less, a principle that was true to their conception of the real numbers.

There would indeed have been an oversight if the Choice-like principles used had not been true to the conception of the real numbers. If the principles had been inconsistent with that conception, we would appropriately identify their use as a mistake. If the principles had been compatible with the conception though not derivative of it, we would appropriately identify the use of the principles without explicit mention as an oversight. But as a matter of fact we appropriately regard the theorems that we now prove using Choice, reconstructing the old proofs, as true of the real numbers—the very same real numbers that were under investigation all along. Choice is licensed by the idea of arbitrarily picking the members of a set. Its eventual use was therefore perhaps inevitable given the notion that a function is an arbitrary succession of values not subject to a common law. The subject matter was not changed by the use of Choice principles—it had already been changed by Fourier, who gave that characterization of a function.

The leading role, I take it, of the claim of self-evidence—though admittedly I cannot show that this is what Zermelo had in mind—was to establish that to use Choice is to go on with the exploration of the same old real numbers and the same old mathematical subject—analysis—as before. It is not to take up the investigation of something new or more special.

The role of self-evidence is not, or at least not always, to ensure *a priori* truth, and it is not independent of the goal of systematization. Historical

[6] See [Lav92, pp. 325–326] for a discussion of why the generalization would be unwarranted on the basis of a theory of what warrants generalizations that was advanced by Maddy [Mad90], a theory that is based on the notion of a natural kind.

explanations involving self-evidence are compatible with, and indeed an important part of, an understanding of the growth of mathematical knowledge, which takes place along the lines characterized by Kitcher's mathematical naturalism [Kit88, p. 295]. Taking the phenomenon of self-evidence seriously is compatible with the idea that mathematical knowledge evolves in a manner much like that in which scientific knowledge evolves—through the refinement and extension of theories, in a broad sense of the term.

To show that the axiom is extrinsically required, "necessary for science," Zermelo provided a list of important results that are intimately bound up with it, a list that included the basic results of cardinal arithmetic and also results important to analysis and algebra. Since Zermelo's time, that list has become far longer.[7]

Zermelo did not reply in any further detail to Borel, Lebesgue, and Baire, who had argued against the axiom and therefore had stated why they did not find it self-evident. One defense available to him was that Borel, Lebesgue, and Baire *did* despite their protestations find the axiom self-evident: each of them had used it—or rather various equivalent principles or consequences of it—in their work without question [Moo82, §1.7]. That work formed part of Zermelo's positive argument for the self-evidence of the axiom.

Perhaps Zermelo thought it unnecessary to reply to Borel, Lebesgue, and Baire: their objections to his work had given rise to a correspondence with Jacques Hadamard [BBHL05], who had ably defended Zermelo's axiom. Why did Borel, Lebesgue, and Baire doubt the axiom? Borel [BBHL05, p. 273] required that a theorem, to be "completely irreproachable," be "a precise result expressible in a finite number of words," and he said that the paradoxes arise "because sets that are not really defined are introduced." Lebesgue said "that to define a set M is to name a property P which is possessed by certain elements of a previously defined set N and which characterizes, by definition, the elements of M" and that

> The question comes down to this, which is hardly new: *Can one prove the existence of a mathematical object without defining it?*
>
> This is obviously a matter of convention. Nevertheless, I believe that we can only build solidly *by granting that it is impossible to demonstrate the existence of an object without defining it.* [BBHL05, p. 265].

And Baire expressed the opinion [BBHL05, pp. 263–264] that the infinite is "in the realm of *potentiality*" and so infinite objects are merely given or defined by convention. To go further "the meaning of these words [set, well-ordered set] must be extended in an extraordinary way and, I would add, a fallacious one." In each case, and indeed in the case of every mathematician of whom I am aware who has expressed reasons for doubting the axiom in print, the objection presupposes that every set is somehow associated with a definition of some kind. The objection is then simply that the axiom provides no means of *defining* a choice function. The great difficulty in defining such functions in standard cases, like the set of all nonempty subsets of the real numbers, suggested that supplying a definition would not be possible in general.

Hadamard's reply was straightforward:

> What is certain is that Zermelo provides no method to carry out *effectively* the operation which he mentions, and it remains doubtful that anyone will be able to supply such a method in the future. Undoubtedly, it would have been more interesting to resolve the problem in this manner. But the question posed in this way (the effective determination of the desired correspondence) is nonetheless completely distinct from the one that we are examining (does such a correspondence exist?). Between them lies all the difference, and it is fundamental, separating what Tannery calls a *correspondence* that can be *defined* from a correspondence that can be *described.* [BBHL05, p. 262]

Whatever notion of set Zermelo and Hadamard were employing, it was expressly not one that required that every set be associated with a definition.

We may now briefly characterize the situation, following Maddy [Mad90, pp. 121–123] in important respects, in these terms: The opponents of the Axiom of Choice were employing the logical notion of collection. The Axiom of Choice is, at best, dubious for logical collections, and it is certainly not self-evident or otherwise suitable for adoption as a basic principle concerning such collections. The supporters of the Axiom of Choice were employing a quite different notion of collection, a combinatorial notion of the sort that originated with Cantor, that is to say a notion according to which collections consist of members enumerated in a perfectly arbitrary way. Such collections exist independently of our ability to give a defining principle, and the Axiom of Choice is indeed self-evident for them, even though it is dubious for logical collections. Cantor had even taken a close cousin of Choice, the

[illegible] ~ NSC

[7] For a detailed history and discussion of the many theorems proved before and after Zermelo's work that entail some form of the axiom, see [Moo82].

Well-Ordering Principle, to be the basic principle for combinatorial collections. The real disagreement was about whether mathematicians should employ logical collections or combinatorial collections, not about Choice. The verdict of history has been that mathematicians should employ combinatorial collections—and hence Choice—for reasons we discuss later in this section.

Cantor's early theory ran into difficulties because it is not clear that the collection of subcollections of a well-ordered collection can be well-ordered. As a result Cantor was not sure whether the power set of a combinatorial collection is itself a combinatorial collection.

Zermelo's substantial contribution was that he succeeded in making it plausible that the Power Set Axiom is compatible with the combinatorial notion of collection, and that it is in fact an illuminating supplement to that notion. First of all, as Zermelo's theorem shows, in the presence of Power Set (and other simple Cantorian principles), Well-Ordering and Choice—apparently two distinct characteristics of combinatorial collections—become provably equivalent. That is, they merge to become a single characteristic. Second, Zermelo's theorem provides some reason to believe the Well-Ordering Principle for the power sets of enumerative collections: We already believe Choice for at least one such collection, and Well-Ordering follows. That argument is based on the fact that the Axiom of Choice is self-evident for the set of real numbers, which is essentially the set of all subsets of the natural numbers and is thus a power set of an enumerative set. For the set of real numbers, Choice is part of the program of freeing the notion of a function from that of a rule.

There is a third notion of collection: the collections actually employed in mathematics, the *mathematical collections*.[8] I am not sure that deserves to be called a notion, since it is defined historically, but no matter. As Cantor basically discovered late in his career, the infinite mathematical collections can all be generated starting from the set of natural numbers using the power set operation and a few simple methods of combination. By showing that the Power Set Axiom is a plausible supplement to the Cantorian notion of a combinatorial collection, Zermelo made it possible to identify the mathematical collections with the combinatorial ones.[9]

[8] Maddy used the term *mathematical collection* for pretty much what I have called combinatorial collections, and she associated such collections with the modern iterative conception of sets, often attributed to Zermelo, which is discussed in §5 [Mad90, pp. 102–103, 121].

[9] Zermelo's collections are of a different sort from Cantor's. I see the one sort as a natural outgrowth of the other, and so I have used the same term. The issue is, so far as I can see, merely a terminological one.

In Cantor's early theory the Well-Ordering Principle had, in effect, served as the only criterion of sethood. That gave a clear characterization of what sets are. The Axiom of Choice is not suitable to serve as a criterion for sethood, and so Zermelo had to supplement it with an *ad hoc* list of supplementary principles. Each of his axioms is true to the combinatorial notion of collection—they all derive from the Cantorian theory. But Zermelo's arbitrarily selected list of axioms is not suitable for characterizing the notion of a combinatorial set.

The two sides of the debate about the Axiom of Choice were really disagreeing about which type of collection, logical or combinatorial, mathematical collections are. There was little disagreement that the Axiom of Choice is not self-evident, and indeed likely false, of logical collections or that the axiom is indeed self-evident of combinatorial collections. As Zermelo argued, and as subsequent mathematical developments have shown, the Axiom of Choice plays an important role in mathematics, and therefore the right notion is the combinatorial one.[10]

Zermelo had shown that the combinatorial notion of collection could be identified with the mathematical one, but that does not show that the logical notion cannot be identified with the mathematical one in some different way. One would lose the Axiom of Choice in so doing, but in the first decade of the century that was perhaps not yet decisive. Indeed, that seems to have been essentially the attitude of Russell in developing the theory of types (letter to Jourdain, 1905 [GG77, p. 55]): "I don't think the continuum of real numbers is upset by the multiplicative class difficulty. Also I have hopes that much will be discovered to circumscribe the difficulty; for all we have at present is a mere absence of proofs of propositions which are very likely to be true." But there are other substantial difficulties with identifying the logical collections with the mathematical collections, as the earlier description of the theory of types made clear. Logical collections are intimately connected with the incon-

[10] The opponents of the Axiom of Choice, all supporters of the logical conception of collection, did not just argue that the axiom is false of logical collections. Many also argued that the very idea of a combinatorial collection is incoherent—since infinite collections can only be introduced by means of definitions. I have omitted those arguments here. As we have mentioned, Poincaré had ideas along those lines. The most fully worked out position is that of Brouwer. Both Poincaré and Brouwer are discussed in Chapter VI.

sistent Comprehension Principle, and so one must confront the paradoxes in any successful attempt to identify logical with mathematical collections. But natural ways to handle the paradoxes also rule out impredicative definitions, as we saw in the case of the theory of types, where the need to allow such definitions led to the *ad hoc* Axiom of Reducibility. Since ordinary mathematics is impredicative, the use of logical collections is apparently blocked. Much progress had been made in this century in reconstructing parts of mathematics within one or another more or less predicative framework, but the fact remains that the ordinary practice of mathematicians cannot be captured in that way and many theorems cannot be reconstructed.[11] The mathematics of today, and of Zermelo's day, allows impredicative definitions and requires the Axiom of Choice. For those extrinsic reasons, we have come to recognize that ordinary mathematical collections are combinatorial collections. No mathematical theory of logical collections adequate to encompass the mathematical collections is as yet available.[12]

For various extrinsic reasons, the theory of mathematical collections is seen to be the theory of combinatorial collections, and the Axiom of Choice is self-evident for such collections. The controversy that surrounded the Axiom of Choice is usually assumed to cast doubt on the self-evidence of the axiom. But there was never a controversy about the axiom itself, which is in fact an uncontroversially self-evident principle about combinatorial collections. The controversy was over two notions of collection, logical and combinatorial, and the combinatorial notion has apparently won. As a modern set theorist, Donald A. Martin, put it,

> much of the traditional concern about the axiom of choice is probably based on a confusion between sets and definable properties . . . Once this kind of confusion is avoided, the axiom of choice appears as one of the least problematic of the set theoretic axioms. [Mad90, p. 124]

The Axiom of Choice is *not,* as it is often taken to be, an example that shows that mathematicians disagree about "self-evidence," an example that casts doubt on that notion. On the contrary, the axiom is just one more example of the surprisingly broad agreement among mathematicians about what is, and what is not, self-evident, and of the real role such considerations play in mathematics.

[11] For more information, see, for example, [Fef77].

[12] Since the extensions of properties are logical collections, to the extent that they are taken to be collections constituted of members at all and not something more Fregean, the considerations in the text cast doubt on their mathematical utility and even their cogency. At the very least, the support that the use of such extensions is often thought to gain from the successful mathematical theory of sets is in fact strikingly absent.

As we have seen, the notion of combinatorial collection is in need of clarification. Logical collections need only be considered in an incidental way henceforth, since logical collections have never played more than an incidental role in mathematics in general and in set theory in particular. The paradoxes, which have often been thought to pose the central problem in making sense of the theory of sets, are, along with the logical collections for which they arise, a side issue. The real issue is how infinite collections can have combinatorial properties. Cantor's private answer, that God can manipulate them much as we manipulate finite collections, may not seem very helpful, but it is our historical starting point.

To see how set theory and our understanding of it have developed since 1908, it is necessary to look at Zermelo's axioms. Much of what has happened was in reaction to them. Here they are:

The first is Axiom IV.2.6, the Axiom of Extensionality. The second axiom is closely related to the Cantorian axioms IV.2.2 and IV.2.3.

AXIOM 1.1 (ELEMENTARY SETS). *There is an empty set. For any object a, there is a set $\{a\}$ with a as its only member. For any two objects a and b, there is a set $\{a, b\}$ that contains them and nothing else.*

AXIOM 1.2 (SEPARATION). *"Whenever the propositional function $\mathfrak{E}(x)$[13] is definite for all elements of a set M, M possesses a subset $M_{\mathfrak{E}}$ containing as elements precisely those elements x of M for which $\mathfrak{E}(x)$ is true." [Zer08b, p. 202]*

The Separation Axiom relies on the following definition [Zer08b, p. 201]:

> A question or assertion $\mathfrak{E}$ is said to be *definite* if the fundamental relations of the domain, by means of the axioms and the universally valid laws of logic, determine without arbitrariness whether it holds or not. Likewise, a 'propositional function' $\mathfrak{E}(x)$, in which the variable term x

[13] The symbol is a German capital *E*.

ranges over all individuals of a class $\mathfrak{K}$,[14] is said to be *definite* if it is definite for *each single* individual x of the class $\mathfrak{K}$.

The Separation Axiom is immediate from Cantorian Definition IV.2.5. At least, that is true for a suitable clarification of "definiteness." The notion of definiteness was the main source of controversy concerning Zermelo's axioms, and we shall discuss it in detail in §3.

Axiom 1.3 (Power Set). *Every set T has a power set, that is, a set that contains exactly the subsets of T.*

As we have discussed above, the Power Set Axiom does not follow from the Cantorian Axioms. It was, in fact, their downfall.

Axiom 1.4 (Union). *Every set T has a union set, that is a set that contains exactly the members of members of T.*

The Union Axiom follows from Cantorian Definition IV.2.5 and Axioms IV.2.7 and IV.2.10.

Axiom 1.5 (Choice). *If T is a set of pairwise disjoint nonempty sets, then there is a subset of the union of T that has exactly one member in common with each member of T.*

The Axiom of Choice follows from Cantorian Axiom IV.2.10.

Axiom 1.6 (Infinity). *There is a set that contains the empty set and the set $\{a\}$ for each of its members a.*

The Axiom of Infinity follows from Cantorian Axiom IV.2.4 and Definition IV.2.5.

I have mentioned how most of Zermelo's axioms can be obtained from the Cantorian axioms. The other direction is more subtle, since Cantorian Definition IV.2.5 does not follow from Zermelo's axioms. That is partially why I said that Zermelo's axioms were not, as he claimed, "the principles required for establishing the foundations of" set theory. That is the story of the next section.

[14] The symbol is a German capital K.

§2. The Axiom of Replacement

Fraenkel [Fra21, p. 97] and Skolem [Sko23b, p. 296] discovered that, in Skolem's words (Fraenkel's are similar), "Zermelo's axiom system is not sufficient to provide a complete foundation for the usual theory of sets." Each had noted that, if Z_0 is Zermelo's official counterpart for the natural numbers within his theory of sets and $\mathcal{P}$ is the power set operation, which takes each set to the set of all of its subsets, then one cannot prove in Zermelo's system that the set $\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$ exists. Skolem gave a proof.[15] Without that set, as Fraenkel observed [Fra21, p. 97], one cannot prove the existence of $\aleph_\omega$.

Fraenkel [Fra22b, p. 231] and Skolem [Sko23b, p. 297] independently proposed the same remedy for the inadequacy in Zermelo's system, namely, introducing a new axiom, which Fraenkel named the Axiom of Replacement:[16]

Axiom 2.1 (Replacement). *The range of a function on a set is itself a set.*[17]

Fraenkel observed [Fra21, p. 97] that it would have sufficed for the problem at hand to introduce an extended axiom of infinity, one that asserted that

$$\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$$

exists. But if Zermelo's system were to be extended in that way, the resulting system could be shown inadequate in much the same way as before, and so a more general principle was required [Fra22b, p. 231].

The axiom solves the problem of ensuring the existence of

$$\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$$

since there is a function that takes any natural number n (actually, its counterpart in Z_0) to the result $\mathcal{P}^n(Z_0)$ of applying the power set operation to Z_0 n times. The range of that function on Z_0 is the required set. That

[15] He observed that $V_{\omega+\omega}$ is a model of Zermelo's axioms that does not contain the set. See §4 for the definition of $V_{\omega+\omega}$.

[16] Both [Moo82] and especially [Hal84] provide useful histories of the Axiom of Replacement.

[17] Skolem allowed partial functions. The two versions are equivalent.

proof mimics our intuitive reason for believing that if Z_0 exists then so does $\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$: we just replace each member of Z_0, that is, each natural number n, with the corresponding member of the set, $\mathcal{P}^n(Z_0)$. Neither Fraenkel nor Skolem doubted the existence of the set, and they independently arrived at the same method of showing that it exists.

The axiom as stated is ambiguous since it is not clear what functions are to be allowed. Fraenkel did not publish an answer to that question until later, and so his version of Replacement was, initially, ambiguous. Skolem gave a precise answer. For both Fraenkel and Skolem the functions to be allowed are the "definite" functions. An answer to the question is thus a solution to the problem of making sense of Zermelo's notion of "definite." That is the topic of the next section.

Replacement is self-evident for combinatorial collections and an immediate consequence of Cantor's theory: If we form one collection by replacing the members of another, a well-ordering of the new collection is determined by a well-ordering of the original one. More formally, if we enumerate the members of S using F, and if we pick the members of a new collection by using f on S, then $f \circ F$ (with any duplications deleted) enumerates the new collection, which is the range of f on S, showing that range to be a combinatorial collection.

Replacement has various precursors. Cantor did not give any argument for Replacement, but he did state something like it [Can32a, p. 114], though as a truth not a postulate: "Two equivalent multiplicities are both 'sets' or are both inconsistent." That has as an immediate consequence that the range of a one-to-one function on a set is a set, which is a version of Replacement. (Though it is a special case, full Replacement follows from it using Zermelo's axioms.) Dimitry Mirimanoff [Mir17a, p. 49] stated the Axiom as a "Postulate," apparently because it was required to develop his theory of set-theoretic representatives of ordinal numbers.[18] That motivation does not have the clear intuitive character of the motivation of Fraenkel and Skolem, and the postulate was not part of a complete axiomatization of set theory.

Neither Fraenkel nor Skolem advocated adding the Axiom of Replacement to Zermelo's system. Neither investigated its consequences. Skolem just said that "we could introduce" the axiom [Sko23b, p. 297]. But Skolem had general doubts about the utility of axiomatic set theory. (See §3.)

[18] His representatives are essentially what are today generally known as "von Neumann ordinals," which will be described later in this section. The only difference is that Mirimanoff allowed urelements but apparently not an empty set and so he identified the number 0 with an arbitrary fixed urelement e instead of with the empty set.

Fraenkel in some ways opposed the Axiom of Replacement. (See [Hal84, pp. 296–297] for a summary.) The sets produced using Replacement are, according to Fraenkel, very large, while those that lead to paradoxes are too large. That suggested to Hallett that Fraenkel was "suspicious" of the axiom [Hal84, p. 296], presumably suspicious that it might lead to contradictions. But I do not think so, and to the best of my knowledge Fraenkel never said so.

Fraenkel's doubts seem to me to lie in another direction, since he said forthrightly that Zermelo's axioms "are not sufficient for the foundation of legitimate set theory" ([Fra21, p. 97], translation [Hal84, p. 280]), and he never seemed to doubt the existence of $\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$. Moreover, he acknowledged, after the work of John von Neumann to be described next, that the axiom is necessary for the theory of ordinal numbers [Fra25, p. 251]. As Hallett put it, "What he seems to challenge is that any of this extra content which the axiom furnishes is set-theoretically important" [Hal84, p. 297]. (See, for example, [Fra25, pp. 251–252] or [Fra28, p. 310].)

Fraenkel believed that "general set theory" does not need the Axiom of Replacement, although some special results, including the theory of ordinal numbers, require it. That belief was entirely reasonable in the 1920s: Ordinary mathematics, including Fraenkel's general set theory, is as a matter of fact concerned exclusively with objects that have counterparts in the set $\bigcup\{Z_0, \mathcal{P}(Z_0), \mathcal{P}(\mathcal{P}(Z_0)), \ldots\}$. That set contains counterparts of the real and complex numbers, functions from real numbers to real numbers, function spaces, and so forth. All of those sets can therefore be proved to exist in Zermelo's theory. No consequence of the Axiom of Replacement that is a part of ordinary mathematics—that is, no consequence that could even be stated without making use of sets going beyond those—was discovered until 1975.[19] To be sure, other theorems may have been proved using Replacement, but they could perfectly well have been proved without it.[20]

[19] The first such result is that every Borel game is determined. See [Mar75] for the relevant definitions and the proof of the result, and see [Fri71] for the proof that Replacement is required and the claim, made prior to 1975, that Replacement plays no role in ordinary mathematics. Further theorems of ordinary mathematics that cannot be proved without Replacement may be found in [HMŠS85].

[20] Though modern mathematics is permeated with set theory, and set theory orig-

Fraenkel's hesitation to add Replacement to Zermelo's system had to do with doubts about the utility of Replacement, not with doubts about its truth. The fact that Replacement had no known applications outside of the theory of infinite ordinal and cardinal numbers but was nonetheless accepted as true shows that, whatever the basis on which Replacement was seen to be true, that basis must have been one distinctively concerned with the infinite.

In 1923 von Neumann worked out important consequences of Replacement [vN23], an axiom that he said [vN23, p. 347] "fills a substantial gap in Zermelo's axiomatization." When W is a well-ordered set, he defined a *numeration* of W to be a function f such that for all w in W $f(w) = \{f(u) : u < w\}$, that is, such that $f(w)$ is the range of f on the predecessors of w. The need to rely on the Axiom of Replacement to show that well-ordered sets have numerations is clear. He defined the range of f on W to be an *ordinal number* of the well-ordered set W. He then showed, still making strong use of the Axiom of Replacement, that every well-ordered set has a unique ordinal number and that similar well-ordered sets have the same ordinal number, which shows that the use of the term *ordinal number* is legitimate; he characterized the ordinal numbers in several different ways; he showed that the ordinal numbers are well-ordered by the membership relation; and he justified definition by induction on the ordinal numbers, which made it possible to introduce addition, multiplication, and exponentiation of ordinal numbers directly, without the use of auxiliary set-theoretic notions like that of ordered sets.

Once definition by induction on the ordinal numbers has been justified, it becomes possible to describe the von Neumann ordinal numbers by saying that each one is the set of its predecessors. Thus 0 is $\varnothing$, 1 is $\{\varnothing\}$, 2 is $\{\varnothing, \{\varnothing\}\}$, and so forth. Cantor had treated the ordinal numbers as separate objects that were obtained from well-ordered sets by "abstraction." Zermelo had axiomatized set theory without ordinal numbers. But von Neumann had shown how to introduce ordinal numbers as sets, making it possible to use them without leaving the domain of sets. The Axiom of Replacement is crucial. It is used to show that every well-ordered set has an ordinal number and to justify definitions by induction on the ordinal numbers.

Though von Neumann apparently did not know it, Mirimanoff had already shown using Replacement that to every ordinal, and hence to every well-ordered set, there corresponds a unique von Neumann ordinal number and that similar well-ordered sets have the same von Neumann ordinal number; he had characterized the von Neumann ordinal numbers in several different ways; and he had shown that the von Neumann ordinal numbers are well-ordered by the membership relation [Mir17a, Mir17b]. He therefore, it seems to me, deserves some of the credit, and so I shall henceforth refer to the so-called von Neumann ordinal numbers as Mirimanoff–von Neumann ordinal numbers. But Mirimanoff did not propose identifying the ordinal numbers with the Mirimanoff–von Neumann ordinal numbers. He did not justify induction on the Mirimanoff–von Neumann ordinal numbers; indeed he made implicit use of induction on well-ordered sets to introduce them [Mir17a, p. 45]. He used Replacement to show that to every ordinal number—a notion taken to be antecedently understood—corresponds a Mirimanoff–von Neumann ordinal number [Mir17a, p. 49], instead of showing what his techniques sufficed to show, that to every well-ordered set corresponds a Mirimanoff–von Neumann ordinal number. He therefore had not shown that one could introduce Mirimanoff–von Neumann ordinal numbers initially as the sole notion of ordinal number. He had only shown that they could be used as a substitute for the ordinal numbers after the ordinal numbers had been used to introduce them.

§3. Definiteness and Skolem's Paradox

This section begins with a brief expository aside.

A *first-order* logic is one in which the quantifiers range only over the members of a domain. A *second-order logic* is one in which there are also quantifiers that range over things that determine relations and operations on a domain. Leopold Löwenheim [Löw15] usually receives credit for the distinction. See [Moo88a] for a careful history.

Second-order quantifiers might, for example, range over relations on a domain, operations on a domain, propositional functions on a domain, or collections of members of a domain. For simplicity, I shall just consider second-order quantification over relations.

The distinctive feature of second-order logic is that it presupposes that,

inated with the ordinal numbers, the ordinal numbers are rarely required outside of set theory, as just discussed in the text. Theorems concerning, for example, Cantor's derived sets seem to involve the ordinal numbers and hence Replacement, but they are a part of ordinary mathematics. The most convenient and natural formulations of such theorems do make use of the ordinal numbers, but they can be reformulated to avoid them and Replacement.

given a domain, there is a fact of the matter about what the relations on it are, so that the range of the second-order quantifiers is fixed as soon as the domain is fixed. I shall not, in the end, make any essential use of second-order logic, and so I shall not endorse the presupposition. Nonetheless, it is important to see that the presupposition is a part of the use of second-order logic for many foundational purposes. There are a number of logical systems in which there are quantifiers that are second-order in that they range over something like relations over a domain, but in which the system is determined in an additional way—by giving a set of relations in addition to a domain or by axioms concerning the quantifiers. For present purposes—that is, with respect to Skolem's paradox—such systems may as well be first-order, and we do not include them in second-order logic. See [Sha91, Chapter 4] for an elementary exposition.

Most authors call all of the systems of logic with quantifiers over relations second-order—including those systems I just excluded. That is the natural thing to do if, for example, you think that a logic should be specified by the rules of its use rather than by truth conditions given in terms of domains and relations on domains. You might think that a logic should be specified by the rules of its use, for example, because it is not clear how domains and relations can be specified short of by giving the rules of reasoning about them. The logic that I am simply calling second-order is usually called "full" or "standard" second-order logic.

I do not wish to take a stand on the issue of which is the appropriate way to specify a logic. After all, I am arguing that logic has much less to do with set theory than is ordinarily supposed, and so the issue is not important here. But it is only full second-order logic that is relevant to Skolem's paradox and to Zermelo's work, and it is therefore a terminological convenience not to have to specify "full" or "standard" every time second-order logic is discussed.

Now, back to our history. Zermelo's second proof of the Well-Ordering Principle made use of a reduction of well-orderings to sets. That is what enabled him to carry out the proof on the basis of an axiomatization of a theory of sets and sets alone. But Zermelo's 1908 axiomatization in fact involved another sort of entity as well—definite propositional functions, which appeared in the statement of the Separation Axiom. That notion was criticized for lack of clarity by many. (See [Moo82, p. 260] for a list.)

In 1910 Hermann Weyl suggested that a property is definite if it can be obtained from =, ∈, and members of the domain using a finite number of definition principles [Wey10, p. 304]. By 1917, he had arrived at a satisfactory

list of definition principles: negation, identification of variables, conjunction, disjunction, substitution of constants, and existential quantification over the domain [Wey18, pp. 4–6, 36]. In modern terms, his idea was that a property is definite if it is definable in first-order logic with parameters.[21] That definition of definite presupposes the notion of finite iteration (of the definition principles) and hence the natural numbers. Thus, Weyl thought that the effort to found the natural numbers on set theory was misguided [Wey18, pp. 6–37].

The main purpose of the book in which Weyl discussed definiteness was the reconstruction of a portion of analysis in a theory that permitted quantification only over the natural numbers. The motivation was to give a predicative reconstruction of a portion of analysis: since real numbers are defined as sets of rational numbers, in effect as sets of natural numbers, a definition of a real number that involves quantification over real numbers is essentially a definition of a set of natural numbers in terms of (quantification over) the set of sets of natural numbers, and hence impredicative. To adopt a predicative program is to give up on standard set-theoretic analysis, and hence to give up on set theory. Perhaps that is why Weyl's suggestion had little impact on the development of set theory.[22]

In 1922 Skolem independently arrived at the same definition of definite as had Weyl [Sko23b, pp. 292–293]. The definite propositional functions of the Separation Axiom were, according to Skolem, just formulas of first-order logic with parameters. The axiom became a schema. The "functions" in Skolem's version of the Replacement Axiom correspondingly were functions definable in first-order logic with parameters. Thus, the Replacement Axiom is also a schema. Let me just give the details for Replacement. Separation is similar but simpler. When $\phi(x, y, u_1, \ldots, u_n)$ is a formula of the language of set theory, there is a corresponding instance of Replacement, which reads:

> Fix sets $a_1, \ldots, a_n$. For any set S, if for every x in S there is a unique y such that $\phi(x, y, a_1, \ldots, a_n)$ (that is, if ϕ with parameters $a_1, \ldots, a_n$ defines a function on S), then there is a set T such that for all y, y is in

[21] According to [Moo88a, p. 135], what Weyl had in mind was closer to *ω logic:* first-order logic with a built-in predicate for the natural numbers.

[22] In 1946 Weyl gave a very readable brief account, in English, of his views and how they relate to those of others [Wey46, pp. 268–279]. Solomon Feferman is an eloquent modern advocate of predicative mathematics [Fef 88].

> T if and only if there is an x in S such that $\phi(x, y, a_1, \ldots, a_n)$. (That is, T is the range on S of the function defined by ϕ with parameters $a_1, \ldots, a_n$.)

That statement, naturally, is intended only as a relatively readable rendering of the official axiom, which is a sentence of first-order logic.

Like Weyl, Skolem [Sko23b, pp. 300–301] dismissed axiomatic set theory as clearly "not a satisfactory ultimate foundation of mathematics."[23] But, unlike Weyl, he dismissed it on the basis of a mathematical theorem about Zermelo's set theory, namely that if it is consistent, then it has a denumerable model. That means, as Skolem emphasized, that Zermelo's set theory has a model in which all of the "sets," even the supposedly nondenumerable ones like $\mathcal{P}(Z_0)$, are natural numbers. In fact the theorem, known as the Löwenheim–Skolem theorem, first proved by Löwenheim and strengthened and given a simpler proof by Skolem, shows that *any* finite or denumerable set of sentences of first-order logic that has an infinite model has a denumerable model. Thus it applies equally well to Zermelo's axioms with or without Replacement and also to every other first-order theory of sets.

There is no mystery about what has gone wrong: "Skolem's paradox" is not a formal contradiction. The natural number that is "$\mathcal{P}(Z_0)$" in a model of Zermelo's axioms with domain the natural numbers has only denumerably many "members," since there are only denumerably many natural numbers. But the theorem "$\mathcal{P}(Z_0)$ is nondenumerable" when applied in the model says only that there is no number in the model that within the model plays the role of a one-to-one correspondence between the number "$\mathcal{P}(Z_0)$" and the number "Z_0," where I have used scare quotes to emphasize that we are not talking about the real $\mathcal{P}(Z_0)$ and Z_0—if there are any such things—but the numbers that those descriptions pick out in our model with domain the natural numbers. Whatever correspondence we might use to show that "$\mathcal{P}(Z_0)$" is denumerable is just not in the model. If we allow (full) second-order logic, it is trivial to block the paradox: By *fiat* a second-order quantifier $\forall X$ includes every relation on the domain in its range, and hence every set of natural numbers if the natural numbers are in the domain. Thus, the second-order axiom

$$(\forall X)((\forall y)(X(y) \to y \in Z_0) \to (\exists x)(x \in \mathcal{P}(Z_0) \wedge (\forall y)(X(y) \leftrightarrow y \in x))),$$

[23] He later changed his mind. See [Ben85, Geo85].

which says that if X holds only of numbers, then there is an x in $\mathcal{P}(Z_0)$ that is the set of numbers of which X holds, ensures that every set of natural numbers is in $\mathcal{P}(Z_0)$, hence in the domain, which is therefore uncountable. Skolem's paradox is blocked by stipulation. That is not much more helpful than just insisting that we intend our first-order interpretation to be full in the sense that its quantifiers range over all sets[24] and that the membership relation is the membership relation, which blocks the paradox directly.[25]

Skolem went on to point out that the notion "nondenumerable" is unavoidably "relative" in that a set that is nondenumerable in one model (for example, the $\mathcal{P}(Z_0)$ of the denumerable model above, which is nondenumerable in that model) may turn out to be denumerable in another (the model in which we carried out the construction of the denumerable model). He declared that "finite," "infinite," and other notions are similarly relative.[26] He concluded [Sko23b, p. 296] that on any consistent axiomatic basis the theorems of set theory "hold in a merely *verbal* sense." However [Sko23b, p. 300], "many mathematicians—indeed, I believe, most of them . . . do not have an axiomatic conception of set theory at all. They think of sets as given by specification of arbitrary collections."[27] One could, Skolem noted, introduce ab-

[24] In fact all we need require is that the quantifiers range over a union of V_αs. See §4.

[25] One might think it more natural to suppose that we know what the relations on a domain are given the domain than to suppose that we just know what sets there are, which makes second-order logic seem a bit more natural than the "full first-order set theory" just suggested in the text. But the point remains that the two block the paradox in essentially the same way.

[26] Both he and, as we shall discuss below, von Neumann suspected that finitude would be relative like the other notions, but that did not follow from Skolem's results. It does, however, follow from Kurt Gödel's incompleteness theorem, announced in 1930 [Göd30] (using his 1929 completeness theorem too), as Gödel essentially pointed out in a review [Göd34] published in 1934 of [Sko33]. Skolem may not have recognized that, since he claimed to have derived that relativity in [Sko34] using essentially the techniques of [Sko33]. In fact, the result follows easily from Gödel's completeness theorem alone, but that does not seem to have been noticed until 1947 ([Hen47], see also [Kle88, p. 49]). Anatolii Ivanovich Maltsev gave arguments for the result similar to those based on the completeness theorem in 1936 and 1941 ([Mal36, Mal41], compare Robert L. Vaught's remarks [Vau86, p. 377].)

[27] The quote is taken out of context—Skolem was making the point that the source of doubt about the Axiom of Choice is the "demand that every set be definable." But he did intend the point I am using the quote to make in the text, even if it was only in passing.

solutely nondenumerable collections on the basis of nondenumerably many axioms or on the basis of an axiom that yields nondenumerably many first-order consequences. But any such method would be circular.

Skolem raised two additional objections to the use of axiomatic set theory as a foundation for mathematics [Sko23b, p. 299]. First, in order to show any axiom system for set theory consistent, one must presuppose, outside of that set theory, a notion of a proof based on the axioms. But such a proof consists of "an *arbitrary finite number* of applications of the axioms." Thus, "the idea of the *arbitrary finite* is essential"; it must be presupposed, not introduced from within the axiomatic set theory. Skolem might have added, though he did not, that introducing the notion from within an axiomatic set theory to prove that theory consistent is not only circular, but that it is inadequate because the notion of arbitrary finite within the axiomatic system is a relative notion, while the one required of genuine proofs is not. In fact, as Gödel showed in 1930 [Göd30], if Zermelo's theory is consistent, then so is that theory plus a new axiom that says essentially that that very theory is *inconsistent*. Once more, there is no paradox. Models of the strange theory with the new axiom have a nonstandard notion of finite and hence a nonstandard notion of proof. The "proofs" of "inconsistencies in Zermelo's theory" that appear in such a model are not proofs in the ordinary, absolute sense: assuming Zermelo's theory is consistent, they are either infinite or not well founded (that is, they involve loops or infinite descending chains), as is easily seen from outside the model.

Skolem's remaining objection to taking axiomatic set theory as a foundation for mathematics was that it is absurd to define the natural numbers and then prove the induction principle on the basis of axiomatic set theory, since the natural numbers and induction are so much simpler and less open to question than any axiomatic set theory. That objection only shows that one cannot use a set-theoretic basis to justify the natural numbers or to render our theory of them more certain. It does not show that the theory of the natural numbers cannot be absorbed into set theory in a more technical sense. Indeed it can: the theory of the finite Mirimanoff–von Neumann ordinals, for example, provides a perfectly good mathematical substitute for the theory of the natural numbers. Note (though Skolem did not do so explicitly) that any axiomatic theory of the natural numbers will be subject to the same kinds of objections that axiomatic set theories are: the relativity of the notion of natural number, the need to presuppose a notion of finite to prove anything about the axiom system, and so forth.

We now turn to Fraenkel's theory of definiteness, published in 1925 [Fra25, p. 254]. Fraenkel simultaneously defined definite properties and a certain class of functions from the domain of all sets to sets, the Fraenkel functions, by induction.[28] Here is a slightly cleaned up version of the definition:

> The base Fraenkel functions are the power set operation, the union operation, and the constant functions. The Fraenkel functions are closed under composition. When f and g are Fraenkel functions then so is the function of x that takes x to the set $\{f(x), g(x)\}$. When f and g are Fraenkel functions, the propositional functions $f(y) = g(y)$, $f(y) \neq g(y)$, $f(y) \in g(y)$, and $f(y) \notin g(y)$ are definite. When ϕ is a definite propositional function, the function that takes any x to $\{y \in x : \phi(y)\}$ is a Fraenkel function.

Fraenkel felt that his version of definiteness was superior to Skolem's since it did not require considerations of logic and stayed close to Zermelo's version [Fra25, p. 251]. There is, perhaps, some truth in that, since the Fraenkel functions are precisely the ones licensed by Zermelo's axioms. That is, Zermelo's axioms are logically equivalent to Extensionality, Choice, and Infinity, which do not assert that functions exist on the domain of all sets, plus the claim that the domain is closed under the Fraenkel functions. Fraenkel's version of Separation says: if m is a set and ϕ is a definite propositional function, then $\{y \in m : \phi(y)\}$ (the set of y in m such that ϕ holds of y) is a set. Fraenkel did not assume any Replacement Axiom.[29] Fraenkel showed [Fra25, Fra26,

[28] The term *Fraenkel function* is due to Hallett [Hal84, p. 283].

[29] One corresponding Replacement Axiom would now be: if m is a set and f is a Fraenkel function, then $\{f(y) : y \in m\}$ is a set. That version is suggested by [Fra26, p. 134]. Von Neumann showed in 1928 that that version of Replacement is derivable from Fraenkel's other axioms [vN28, p. 324]. It therefore does not do the required job. There is a different version of Replacement that is more in line with Fraenkel's procedure: since the Replacement Axiom asserts the existence of a function, that function should be added to the definition of Fraenkel functions. Define Fraenkel* functions by adding the following condition to the definition of Fraenkel functions: If f is a Fraenkel* function, then so is the function that takes x to $\{f(y) : y \in x\}$. Fraenkel did at one time suggest that adding Replacement would require widening the notion of Fraenkel function [Fra25, p. 271]. Unfortunately, it is a consequence of von Neumann's result that the new version is no better than the old one. Indeed, it follows easily from von Neumann's work that the Fraenkel* and Fraenkel functions coincide. Von Neumann [vN28, p. 323] suggested a different addition to the definition of Fraenkel function that does the job: Let $\phi(x, y)$ be

Fra32] that his version of Zermelo's system sufficed for the development of much of set theory. As we noted in the previous section, he explicitly admitted [Fra25, p. 251] that von Neumann [vN23] had shown that the system without Replacement does not suffice for the "special" theory of ordinals and cardinals.

In 1925, von Neumann published an axiomatization of set theory [vN25]. He had actually developed it two or three years earlier [Hal84, p. 283]. He credited Fraenkel with the Replacement Axiom [vN25, p. 398], ignoring Skolem's contribution, though he cited Skolem's paper in which Replacement appeared [Sko23b] concerning Skolem's paradox.

Von Neumann axiomatized, in effect,[30] a theory of what he called *classes*. Among the classes, certain ones are members of other classes. Such classes are *sets*. Thus, much like Cantor, he allowed the class of all sets, the class of all Mirimanoff–von Neumann ordinal numbers, and so forth. As he knew, such classes would lead to paradoxes if they were allowed to be members of other classes—hence the distinction between sets and classes, which is remarkably similar to Cantor's distinction between sets and absolutely inconsistent multiplicities.

Von Neumann's new system included the following distinctive axiom:

AXIOM 3.1 (LIMITATION OF SIZE). *A class is of the same power as the universe of sets if and only if it is not a set.*

To say that a class S is of the same power as the universe of sets means that there is a function from the class onto the universe of sets, that is, a class F of ordered pairs such that each member of the class S is the first component of exactly one pair in the class F and such that every set is the second component of at least one pair in the class F. In effect, the axiom says that the only way a class can fail to be a set is by being as large as possible—as large as the class of all sets. Von Neumann argued for that by saying [vN25, p. 402] that it clarified existing confusion, is extraordinarily powerful, and "enlarges rather than restricts the domain of set theory." That limitation of size picture is extremely close to Cantor's, discussed in §IV.2.[31] The Axiom of Replacement—the range of a function on a set is a set—is immediate from von Neumann's axiom, since the range can be no larger than the original set. Cantor's argument from the letter to Dedekind shows that the class of ordinal numbers is not a set. Hence, it follows from von Neumann's axiom that there is a function from the class of ordinal numbers onto the class of all sets. Thus, the class of sets can be well-ordered, and so Choice follows from the axiom as well.

Technical Remark. In fact, a strong form of Choice follows from Limitation of Size: it says that for every class S there is a class T such that for every x, if there is an ordered pair in S whose first element is x, then there is exactly one such pair in T. Von Neumann showed [vN29, pp. 506–508] that the system with Limitation of Size is consistent if and only if the system with the strong form of Choice just introduced plus Replacement is consistent. He also showed by assuming Foundation (see §4) plus his other axioms that Limitation of Size is equivalent to those two axioms. (The consistency result is immediate from the other result plus his theorem [vN29, pp. 494–508] that the system obtained from his by replacing Limitation of Size by Strong Choice and Replacement is consistent without Foundation if and only if it is consistent when Foundation is added.) Since we shall consider Limitation of Size without Foundation in subsequent sections, let me mention that it is straightforward to check that all von Neumann's proof requires is that there be a function from the ordinals to sets such that every set is a member of some member of the range. The equivalence therefore also holds in models of the Anti-Foundation Axiom (see §4).

In von Neumann's system, the Separation Axiom is just the fact that the intersection of a set and a class is a set. Von Neumann thus, in effect, identified

a definite propositional function. If for each x there is a set that has as members exactly those y such that $\phi(x, y)$, then the function that takes each x to the set of y such that $\phi(x, y)$ is a Fraenkel function.

[30] Von Neumann took functions to be primary, not sets. He introduced sets as characteristic functions, that is, as functions from the domain to two objects, functions like the ones Cantor used in his diagonal argument. No one has followed von Neumann in taking functions to be primary, and so I shall ignore that, acting as if he had taken sets to be primary. In the transformed picture, von Neumann's functions become classes of ordered pairs such that every set is the first element of exactly one member of the class.

[31] Von Neumann believed that Cantor's set theory was the "naive" set theory that is in fact due to Russell. He attributed the basic limitation-of-size idea to Zermelo [vN25, p. 397]. In so far as the Separation Axiom only introduces subsets of an already given set, there is a kind of limitation-of-size principle to be found in Zermelo's system—a limitation to subsets—but it is not clear how it relates to von Neumann's proposed limitation on the *cardinality* of sets, a limitation that, as we have seen, results from Cantor's conception. See §5.

the definite propositional functions of Zermelo's Separation Axiom with the classes. To ensure that there are enough classes, von Neumann used axioms that entailed that any collection of sets is a class if it is first-order definable (with class parameters). Since first-order formulas are built up from atomic formulas using, say, negation, disjunction, and existential quantification, it is enough to ensure that the complement of any class is a class, that the union of two classes is a class, that the domain of any binary relation is a class,[32] and so forth. Von Neumann thus obtained a theory in which all of the relevant notions are axiomatized as members of the domain. In contrast, Skolem had required the auxiliary notion of a first-order formula, and Fraenkel had required that of a Fraenkel function. Von Neumann's theory is finitely axiomatized. No schemas are required.

Von Neumann said [vN25, p. 395] his work was in what would today be called a "formalistic" spirit: "one understands by 'set' nothing but an object of which one knows no more and wants to know no more than what follows about it from the postulates." Nonetheless, he carefully observed [vN25, p. 403] that his axioms "are nothing but trivial facts of naive set theory." That observation was important to him because it showed that the axioms, in the specific sense indicated, do not require too much [vN25, p. 403]. Thus, it was important to von Neumann that the axioms be "evident and reasonable," a constraint that is not formalistic. (Elsewhere [vN25, p. 402] he qualified his claim with respect to his main axiom, which he admitted is stronger than "what was up to now regarded as evident and reasonable.")

Von Neumann went on to sharpen some of Skolem's arguments against axiomatic set theory. He noted that probably no theory that has infinite models is *categorical*—that is, is such that all of its models are isomorphic—and so no theory of any infinite mathematical system can characterize that system. He concluded [vN25, p. 412], "This circumstance seems to me to be an argument for intuitionism."[33] He noted that the notion of "well ordering" is subject to Skolem's relativity. About the relativity of the notion of finitude, he said that it is difficult to say whether this militates more strongly against its intuitive character or its set-theoretic formalization. It counts against both, since it shows [vN25, p. 413] that we lack "any foothold that would enable us to make the definition of 'finite' determinate."

[32] A binary relation is a class of ordered pairs. The domain of a binary relation is the class of all x for which there is a y such that the ordered pair $\langle x, y\rangle$ is in the relation.

[33] Intuitionism, a philosophy of mathematics that rejects set-theoretic foundations, is discussed in §VI.2.

When combined with Skolem's arguments, those of von Neumann amount to a devastating criticism of our present-day axiomatic foundations of mathematics on the basis of set theory: Those foundations rely on a notion of proof, which requires a notion of finitude for its definition. But, once we specify the notion of definiteness, our axioms enable us to show that the definition of finitude they provide is an inadequate foundation for the notion of proof. To be sure, every theorem of mathematics has a counterpart within set theory—including the whole theory of finitude, based on the "finite" Mirimanoff–von Neumann ordinals. But that theory cannot serve as a basis for the notion of proof, and hence set theory cannot serve as a basis for an axiomatic mathematics, even if concerns about the certainty of the basis, like consistency or self-evidence, are not at issue. As Skolem said, we cannot make sense of what we are doing without presupposing the notion of "arbitrary finite number." (None of that provides an argument against a realist view of set-theoretic mathematics, on which axioms and proofs play only an incidental role.) The criticism is not directed at the practice of using proofs, a practice that we can certainly acquire without having a theory of proofs or an adequate characterization of finitude. In practice all we need in the way of a theory of finitude is the recognition that any proof we actually encounter in completed form is finite, which is very far from a complete characterization of what it is to be finite. The criticism is directed at a certain kind of attempt to characterize or define what the practice allows in the way of proofs. The attempt fails.

In 1929 Zermelo clarified his own view of definiteness. He began by discussing the various attempts by others at defining definiteness. The difficulty, he said, with eliminating the notion of definiteness in favor of general logic is that there is no widely accepted general logic [Zer29, pp. 339–340]. He clearly had in mind proposals like that of Weyl and Skolem. He criticized Fraenkel for introducing the notion of what we have called Fraenkel functions via a construction, since the construction depends on the notion of the finite numbers, the clarification of which is an important job for set theory [Zer29, p. 340]. He therefore preferred von Neumann's purely axiomatic approach, though he thought that the use of functions made von Neumann's foundation intricate and hard to understand [Zer29, p. 340].

Since we have avoided von Neumann's use of functions, our version of von Neumann's system bears some resemblance to the kind of system Zermelo seems to have had in mind. There are two differences worth emphasizing: Zermelo allowed something like quantification over classes [Zer29, p. 343], and he introduced a restrictive axiom, which said essentially that no proper part of the collection of all classes satisfies the axioms for classes. That axiom was intended to have the effect of ensuring that there were no more classes than those required by the axioms [Zer29, p. 344].

Skolem replied to Zermelo promptly. For Skolem, the question whether a notion of definiteness is presented axiomatically or by means of a construction was merely a matter of formulation [Sko30, pp. 337–338]. He noted that Zermelo's proposal was quite similar to what his own would be, were it to be presented axiomatically.

The main differences between the proposals of Zermelo and Skolem were the restrictive axiom and the quantification over classes. About the restrictive axiom, Skolem asked [Sko30, p. 338]: Since Zermelo did not want to use the notion of finite number, why did he use the notion of proper part? Isn't that also a notion that is to be fixed by set theory? The quantification over classes stood in need of further clarification. Did Zermelo intend to clarify by means of a construction, or further axioms? Most important, if classes were introduced initially without quantification over classes, then allowing quantification over them does not make new sets or classes possible [Sko30, pp. 339–340] and is therefore superfluous. Finally, Skolem noted, much as before, that axioms for set theory will not specify a single model, since they will always have a denumerable model because of the Löwenheim–Skolem theorem [Sko30, p. 340–341].

Zermelo was implicitly relying on what we would today analyze as second-order notions when he quantified over classes and employed parts of the domain. Skolem's request for clarification was a request for a first-order version of those notions, to be supplied by the use of a construction or axioms.

Any first-order theory (that is, any theory formulated in a first-order logic) is subject to Skolem's argument for relativism. Second-order theories are not, but Skolem's request for clarification seems legitimate. Thus, the stand-off between Skolem and Zermelo.[34]

§4. Zermelo

In 1930, Zermelo proposed a new axiomatization of set theory [Zer30, p. 30]: an Axiom of Pairing (any two objects compose a set) much like that of Fraenkel [Fra25, p. 254] replaced his old Axiom of Elementary Sets. The Axiom of Infinity was dropped, on the grounds that it does not belong to general set theory. The Axiom of Choice was not expressly part of the system, since Zermelo thought of it as part of the background logic. The Axiom of Separation took a quite general form: one can separate out of any given set the subset of elements for which a given propositional function holds, where the propositional function can be any whatever, without restriction. As he also put it, any part of a set is itself a set. He mentioned his 1929 article summarized in §3 and Skolem's criticism of it, but simply reserved the right to add to that discussion. He added a strong version of the Axiom of Replacement analogous to his Axiom of Separation and kept the old Axioms of Extensionality, Power Set, and Union. Finally, he added a new axiom [Zer30, p. 31]:

AXIOM 4.1 (FOUNDATION). *Every (descending) chain in which each element is a member of the previous one is of finite length.*

He claimed that that is equivalent to the following: every nonempty part P of the domain has a member that has no members in P.[35] The Axiom of Foundation forbids circles of membership and ungrounded sets [Zer30, pp. 29, 31]. Foundation had a different status from the other axioms, as is indicated by the fact that Zermelo referred to the proposed system as supplemented ZF or ZF′, where Foundation is the supplement. The difference is an important one—Zermelo did not believe that Foundation is true. He commented that the axiom had been satisfied in all useful applications of set theory up to that time, and thus it, provisionally, imposed no essential restriction on the theory [Zer30, p. 31]. He thus apparently believed that while there are non-well-founded sets (that is, sets that are at the top of infinite descending chains) they are of little importance in known applications of set theory. Zermelo used the restriction to well-founded sets to great effect to investigate the models of the supplemented theory.

Zermelo's axiom system is much like so-called ZFC, the axiomatization of set theory that is in most common use today. The two main differences are these: First, he allowed what he called *urelements,* objects that are not sets and have no members, in the domain, while it is conventional to exclude them today. Fraenkel was the first to propose excluding urelements [Fra22b, p. 234], as part of an attempt to give categorical axioms. Skolem mentioned [Sko23b, p. 298] that axioms that do not exclude urelements have

[34] For a modern discussion, and access to the extensive modern literature on the stand-off, see [Sha90]. We return to the subject in §VII.4.

[35] They *are* equivalent given a sufficiently strong form of Choice. Details are omitted, since we never consider dropping Choice in this book.

models both with urelements and without, and he sketched a proof. In much the same way that Zermelo argued for Foundation, Fraenkel argued that urelements serve no mathematical purpose and that eliminating them simplified matters.[36]

Second, Zermelo's Axioms of Separation, Replacement, and Foundation are based on a strong second-order understanding of the notion of definiteness: he considered *all* propositional functions, functions, chains, and parts of the domain, while today we allow only first-order definable (with parameters) propositional functions and functions, and only chains and parts of the domain that are themselves members of the domain. The system Zermelo proposed is essentially what we would today call second-order ZFC, though without Infinity.[37] The first to use a first-order system much like the one so universally adopted today was von Neumann in 1928 [vN28, pp. 321, 323], who named it Zermelo–Fraenkel set theory.[38]

Though the Axiom of Foundation had several precursors, Zermelo did not discuss them. Mirimanoff [Mir17a, p. 42] was the first to distinguish the well-founded sets, which he called "ordinary sets," from the non-well-founded, "extraordinary" ones. Though Mirimanoff never considered the Axiom of Foundation, which would in his terminology have been that all sets are ordinary, he did the next best thing: he restricted his attention to the study exclusively of ordinary sets [Mir17a, p. 39]. That is not far from Zermelo's attitude. After all, Zermelo did not present Foundation as a new truth about sets, as he had presented Choice. He used Foundation to restrict his considerations to well-founded sets, just noting that the other sets have no use. Mirimanoff defined the *rank* of an ordinary set as follows [Mir17a, p. 51]: The rank of an urelement or the empty set is 0, while the rank of any other set is the least ordinal number greater than the ranks of any of its members. He showed, using Replacement, that every ordinary set had, according to that definition, a unique rank.

Skolem [Sko23b, p. 298] mentioned that for any domain that is a model of Zermelo's axioms, the elements of the domain that are not at the beginning of an infinite descending chain also form a model. He was not advocating a restriction, merely noting that Zermelo's 1908 axioms apparently do not determine whether such chains exist.

In 1925 von Neumann [vN25, pp. 404, 411–412] mentioned the possibility of an axiom that there is no function f with domain the natural numbers such that for all n, $f(n+1) \in f(n)$, that is, that there is no infinite descending chain. Zermelo [Zer30] cited that paper for another purpose. Von Neumann did not have his version of Foundation in his basic system. He mentioned it in the second part of the paper, "Investigation of the axioms," primarily to note that it probably did not help to ensure that the theory is categorical, because of the relativity of the notion of descending chain. He did observe [vN25, p. 412] that adding the Axiom of Foundation would not lead to contradictions if there were none already (on the basis of his axioms) and that it would have the "desirable" effect of excluding "superfluous" non-well-founded sets. He published the proof that adding Foundation would not lead to contradictions in 1929 ([vN29, pp. 494–508]; see [Vau85] for an elementary presentation) in a paper in which he also gave (p. 498) the alternative form of the Axiom mentioned by Zermelo and (p. 503) the definition of rank, apparently independently of Mirimanoff.

Zermelo developed the use of rank to understand the structure of the well-founded sets. He went far beyond Mirimanoff. He observed that ranks stratified the well-founded sets: at the zeroth layer is the set of all objects with no members—the empty set[39] and all the urelements. Every *layer* (indexed by an ordinal)[40] is the set of all sets that are made up of objects that occur in any
previous layers. Thus, if we let $V_0(U)$ have as members the empty set plus the members of U, where U is a (possibly empty) set of urelements, and let $V_\alpha(U)$ otherwise be the union of $\bigcup_{\beta<\alpha} V_\beta(U)$ and its power set, then every well-founded set (over the urelements in U) is a member of some $V_\alpha(U)$, and the rank of an object a is just the least α such that a is in $V_\alpha(U)$.[41] The members of each well-founded set lie in preceding layers, and each well-founded

[36] It is ironic that Fraenkel was the first advocate of excluding urelements, since he was also the first to put them to serious mathematical use: in 1922, he proved that Choice is not a consequence of the other axioms if one allows urelements [Fra22a] (assuming the other axioms are consistent). That was not shown without urelements until 1963, by Cohen, see [Jec78, p. 184].

[37] Today one drops Separation, since it is a consequence of Replacement, and one uses a first-order version of Foundation, since the second-order versions follow from that version and second-order Replacement.

[38] The system is Zermelo's, using Fraenkel functions to specify what definiteness means, supplemented by Replacement amended as discussed in a note to §3. Urelements are allowed, and Foundation is not added.

[39] Zermelo took the empty set to be an arbitrarily selected urelement.

[40] Zermelo distinguished between the ordinal numbers and their substitutes in his domains, the Mirimanoff–von Neumann ordinals. I shall suppress the details.

[41] Here and below, I have modernized Zermelo's notation, and I have rephrased his

set serves as material for following ones [Zer30, pp. 29–30]. Zermelo did not emphasize that picture. After all, it did not apply to all of the sets, and he did not view the sets as being in any important sense constructed or built up. He merely said that it helped with his investigation of models of the axioms, and he introduced it as an aid on a par with that of using the Mirimanoff–von Neumann ordinals. The layers are not even mentioned in the final, concluding section of the paper.

Zermelo showed that his axioms serve to guarantee that every model is isomorphic to one of the form $\bigcup_{\beta<\kappa} V_\beta(U)$. Zermelo called models of that form *normal domains*. Normal domains are the subject of Zermelo's investigation. The stratification of sets within each normal domain plays only an auxiliary role. Since every model is isomorphic to a normal domain, we can afford to ignore all other models—results about them will follow immediately from the ones about the normal domains. Moreover, when we disallow urelements and abbreviate $V_\beta(\varnothing)$ (the $V_\beta(U)$ with the empty set of urelements) by V_β, if $\bigcup_{\beta<\kappa} V_\beta$ is a normal domain—a model of Zermelo's axioms—then V_κ is a model of von Neumann's axioms—if we take the sets to be exactly the members of $\bigcup_{\beta<\kappa} V_\beta$. Zermelo showed that a normal domain is characterized up to isomorphism by just two numbers—the cardinality of the set of urelements, which can be any cardinal number, and the least Mirimanoff–von Neumann ordinal number κ not in the model, which can be any strongly inaccessible initial ordinal.[42] Zermelo argued that if there is any normal domain at all that contains infinite members, then strongly inaccessible initial ordinals must certainly exist. Naturally one can assume otherwise, just as one can assume that there are no urelements, but only at the cost of generality [Zer30, pp. 44–45].

Distinct normal domains stack neatly. Suppose, for example, that two nonisomorphic normal domains $\mathfrak{V}$ (German capital V) and $\mathfrak{V}'$ have sets of urelements of the same size. Then they are characterized by distinct ordinals, say κ and κ' with, say, $\kappa < \kappa'$. Zermelo showed that $\mathfrak{V}$ is isomorphic to a substructure of $\mathfrak{V}'$ and indeed that κ is in $\mathfrak{V}'$ and that $\mathfrak{V}$ is isomorphic to the set $\bigcup_{\alpha<\kappa} V_\alpha(U)$ as defined within $\mathfrak{V}'$.[43] That is, in particular, the smaller domain is isomorphic to a set in the larger domain.

results to deal only with the cumulative $V_\alpha(U)$s. He often considered the set of all sets of rank α, that is, in our notation, the set $V_\alpha(U) - \bigcup_{\beta<\alpha} V_\beta(U)$, not $V_\alpha(U)$.

[42] A *strongly inaccessible initial ordinal* κ is an ordinal κ such that κ is of greater cardinality than any of its predecessors, κ is not the least upper bound of a set of cardinality less than that of κ of ordinals that are less than κ, and the power set of any ordinal less than κ has cardinality less than that of κ. One usually requires in addition that κ be greater than ω, but Zermelo allowed ω as a strongly inaccessible initial ordinal since he omitted the Axiom of Infinity from his set theory. With the definition I have given of $V_\alpha(U)$, not every strongly inaccessible initial ordinal gives rise to a model when the set of urelements is large. Zermelo showed that that can be patched up [Zer30, pp. 38–39]. I omit the details.

The above results make strong use of the second-order character of Zermelo's axioms: The use of *all* chains in the Foundation Axiom, instead of just those in the domain, guarantees that the ordinals of a domain really are well-ordered, which a first-order version of the axiom could not have done, and so the ordinals of any model are guaranteed to be (order) isomorphic to "genuine" ordinals. The use of *all* propositional functions in the Separation axiom guarantees that the $V_\alpha(U)$s of different normal domains with the same U are the same when α is in both domains: Suppose not. Then there is a least α in the domains at which the domains differ. Since they differ, one of them must include a set C in $V_\alpha(U)$ that the other omits. But every member of the set C is in both domains, since α is minimal. Separation guarantees that that set is in the other domain, contrary to our assumption, since C is the subclass of $\bigcup_{\beta<\alpha} V_\beta(U)$ determined by the propositional function that holds of the members of C and nothing else. The proof would not go through with Skolem's first-order version of Separation, since the requisite propositional function (x is in C) might not be given by a formula.

Zermelo accepted that there are many normal domains. They are nested as described above. That enabled him to make sense of his use of all propositional functions. Let me speak in terms of the collections that the propositional functions define, the parts of the domain, instead of in terms of the propositional functions themselves. That will serve to put Zermelo's position in the terms of our present-day view, and it will make clear what Zermelo's implicit reply to Skolem was.[44] Zermelo's Separation Axiom says that every part of a set is a set. Skolem's question had been, Isn't the notion of "part" to be fixed by set theory? Zermelo could now reply that it is: A model of set theory is a set in a higher model (that is, a model containing all of the urelements of the first model that has a larger characteristic ordinal), and so the notion of a part of a given model can now be explained set theoretically: a part of the

[43] None of the results requires the Axiom of Infinity. Without it, $V_\omega(U)$, the set of hereditarily finite sets, is a normal domain, and hence included in the analysis. That is probably why Zermelo omitted Infinity.

[44] It is no accident that putting things in terms more congenial to Skolem helps to make contact with today's views. Today's views are direct descendants of those of Skolem.

model is nothing more than a subset of the model, where the term *subset* is used in the sense of a higher model in which the original one is a set. Zermelo assumed that there always is a model higher than any given one [Zer30, p. 46]. The parts of a normal domain, its proper classes, are ordinary sets in a higher domain. Thus was the failure of the axiom system to be categorical turned into a virtue.

Note that all that is required for Zermelo's reply to Skolem is that every model of the axioms is a set in a higher model. If one could show that a similar result held in some system without the Foundation Axiom, then Zermelo's basic philosophical attitude would go over to that system unchanged. That is no idle speculation—it applies to a system of non-well-founded set theory that has some advocates today, ZFC minus the Foundation Axiom plus the "Anti-Foundation Axiom" (ZFC^-+AFA). See [Acz88] for an exact statement of the axiom.[45] Zermelo assumed that there are as many strongly inaccessible initial ordinals as ordinals, and hence that there is a model of set theory for each ordinal. The same result follows for ZFC^-+AFA. Alfred Tarski made a proposal equivalent to that of Zermelo but as an axiom of set theory instead of an assumption about set theory, eight years later [Tar38]. The succession of models also clears up the paradoxes: the proper classes of one model are sets in all higher models [Zer30, pp. 29, 47].

Zermelo went on [Zer32] to publish a strong attack on Skolem's "assumption that all mathematical concepts and theorems must be representable by a *fixed finite system of signs.*" Zermelo thought the resulting relativity of set-theoretic notions ought to convince anyone to abandon Skolem's "prejudice." He went on to say that "our system of signs is always an *incomplete device,* shifting from case to case. It reflects our *finite* understanding of the infinite, which we cannot *immediately* and *intuitively* 'survey' or comprehend, though at least we can approach mastery step by step." (See [Zer32, p. 85] as translated by Gregory H. Moore [Moo80, p. 126].) He went on to propose an infinitary logic, with conjunctions and disjunctions of any set of propositions and well-founded infinitary proofs ([Zer32, pp. 86–88], see also [Zer35]).

Within that logic, arbitrary propositional functions can be used to establish Zermelo's structure theory described above, and, moreover, Zermelo noted that that logic is not subject to the Gödel incompleteness results.

Zermelo summarized his view eloquently in an unpublished manuscript:

> How must a 'domain' of 'sets' and 'urelements' be constituted, to satisfy the 'general' axioms of set theory? Is our axiom system 'categorical,' or does it give a multiplicity of essentially distinct 'set-theoretic models'? Is the idea of a 'set' in contrast to a pure 'class' absolute, determined by logical characteristics or only relative, dependent on the set-theoretic model suitable as a basis at a time? . . .
>
> Every 'normal domain' is a 'closed domain' and therefore can in a higher one also be interpreted as a 'set.' . . . No (closed) normal domain can represent the whole of set theory . . . The whole of set theory is only represented by the 'open' totality of all normal domains. ([Moo80, pp. 131–133], my translation)

He also said that mathematics begins with the infinitary logical assimilation of intuitively given material; it cannot be based on intuition [Moo80, pp. 134, 135]. He did not explain much more. I mention that only because the statement may seem to conflict with the kind of idea of self-evidence that I attributed to Zermelo earlier. It does not. All he means to deny by denying that mathematics is founded on intuition is that mathematics is about space, time, or the like. The notion of assimilation of intuitively given material leaves ample room for self-evidence.

§5. Go Forward, and Faith Will Come to You

There are several important aspects of today's approach to set theory that have not come up in our historical account so far. Chief among them are the idea that axiomatic set theory concerns a single intended domain consisting of all the sets, the iterative conception of sets, and the primacy of the first-order versions of the axioms.

Let me begin by briefly discussing the primacy of first-order versions of the axioms, since there is at least some reason for that in what has come before, even though it has not been emphasized. The second-order quantifie in second-order axiomatizations of set theory essentially range over cl in von Neumann's sense of class—a collection of sets. Since von Ne axioms take the classes to actually be in the domain, they are first o

[45] The construction of a model of ZFC^-+AFA given in [Acz88, Chapter 3] yields a separate model of the axioms for each pure (that is, urelement-free) normal domain. Those models of ZFC^-+AFA are uniquely determined up to isomorphism by Zermelo's parameter (the least ordinal not in the model), and they are nested as required: each model is a set in any higher model, and any higher model contains all the parts of the original model.

the classes in the domain, the temptation to use second-order quantification over them has been removed. Zermelo said (see §4) that he preferred von Neumann's approach to definiteness since it is purely axiomatic, not relying on a construction that is described outside the set-theoretic axioms. He applied that criticism to Fraenkel's proposal, but it applies equally well to the proposal of Weyl and Skolem, which relies on a logic specified outside the set-theoretic axioms. A preference for von Neumann's axioms on whatever grounds is *ipso facto* a preference for first-order axioms.[46]

Skolem's criticisms apparently led Zermelo to change the type of axiom system he favored, but his old reasons for preferring von Neumann's system were adopted by others. In addition, Skolem's insistence that second-order systems required further explanation had the effect of discouraging their use. See [Moo88a] for a discussion of how influences not particularly connected to set theory led to a general preference for first-order logic. Now on to the main business of this section.

Cantor thought of himself as studying sets, not some limited partial domain of sets, and certainly not the formal consequences of axioms. Thus, it is not unreasonable to say that Cantor believed that set theory concerns a single intended model. Since Zermelo just presupposed Cantorian set theory in his 1904 proof of the Well-Ordering Principle from the Axiom of Choice, he may be thought to have taken a similar view—his remarks are so abbreviated that one cannot rule that out.

By 1908, the situation had subtly changed: Since Zermelo argued for the truth of Choice and his other axioms, he was certainly committed to the idea that there are sets, and that they are what is being studied. That also becomes clear in his argument [Zer08a, p. 191] that impredicative definitions are acceptable—since, "after all, an object is not created through such a 'determination.' " On the other hand, in reply to those who had argued that the 1904 proof used principles that let to paradoxes, he said [Zer08a, p. 195], "it is not permissible to treat the extension of every arbitrary notion as a set . . . But if in set theory we confine ourselves to a number of established principles . . . —principles that enable us to form initial sets and to derive new sets from given ones—then all such contradictions can be avoided." He was proposing to abandon the general study of all sets, concentrating on those that could be shown to exist by means of a few principles. Since, as we have seen, his principles were chosen in an *ad hoc* way, there was no reason to suspect that they were strong enough to generate all sets. Indeed they were not strong enough, as the later discovery of the need for Replacement showed.

As we have just seen, Zermelo proposed abandoning the general study of sets when he introduced axioms for set theory in 1908. In the period we have discussed after 1908 no one investigating the foundations of set theory concerned themselves with a single intended model of all sets. Skolem and von Neumann argued against the possibility of determining such a model. Fraenkel followed Zermelo's hint: He abandoned the general study of sets, trying to see instead how few sets were needed to develop the needed theory, rejecting Replacement, introducing an axiom of restriction (which said that no sets existed other than those required by Zermelo's axioms [Fra23, p. 219]), and giving a first-order version of definiteness. He could therefore avoid questions about all sets. Zermelo later (after 1929) did try to clarify what sets are instead of restricting his investigations to those sets whose existence is guaranteed by some principles, but he denied that the sets form a single domain.

Gödel introduced a very different point of view in 1947 [Göd47, vol. 2, p. 180]:

> It might at first seem that the set-theoretical paradoxes would stand in the way of such an undertaking, but closer examination shows that they cause no trouble at all. They are a very serious problem, but not for Cantor's set theory. As far as sets occur and are necessary in mathematics (at least in the mathematics of today, including all of Cantor's set theory), they are sets of integers, or of rational numbers (i.e., of pairs of integers), or of real numbers (i.e., of sets of rational numbers), or of functions of real numbers (i.e., of sets of pairs of real numbers), etc.; when theorems about all sets (or the existence of sets) in general are asserted, they can always be interpreted without any difficulty to mean that they hold for sets of integers as well as for sets of real numbers, etc. (respectively, that there exist either sets of integers, or sets of real numbers, or . . . etc., which have the asserted property). This concept of set, however, according to which a set is anything obtainable from the integers (or some other well-defined objects) by iterated application[a] of the operation "set of",[b] and not something obtained by dividing the totality of all existing things into two categories, has never led to any antinomy whatsoever; that is, the perfectly "naïve" and uncritical working with this concept of set has so far proved completely self-consistent.[c]

[46] Zermelo proposed second-order axioms anyhow, but, as we have seen, Skolem showed that allowing second-order quantification in a system like von Neumann's did not make it possible to define new classes.

> But, furthermore, the axioms underlying the unrestricted use of this concept of set, or, at least, a portion of them which suffices for all mathematical proofs ever produced up to now, have been so precisely formulated in axiomatic set theory[d] . . .
>
> a. This phrase is to be understood so as to include also transfinite iteration, the totality of sets obtained by finite iteration forming again a set and a basis for a further application of the operation "set of".
>
> b. The operation "set of x's" cannot be defined satisfactorily (at least in the present state of knowledge), but only be paraphrased by other expressions involving again the concept of set, such as: "multitude of x's", "combination of any number of x's", "part of the totality of x's"; but as opposed to the concept of set in general (if considered as primitive) we have a clear notion of this operation.
>
> c. It follows at once from this explanation of the term "set" that a set of all sets or other sets of a similar extension cannot exist, since every set obtained in this way immediately gives rise to further application of the operation "set of" and, therefore, to the existence of larger sets.
>
> d. [At this point, Gödel referred in a note to the axiomatization of von Neumann and to those of Paul Bernays and himself, which are based on von Neumann's. All are first order.]

In that brief passage, Gödel introduced the idea that axiomatic set theory is the study of a single intended domain of all sets. He introduced the *iterative conception*—the idea that sets are to be conceived as just the objects obtained by iteration of the "set of" operation, in other words, the idea that sets are to be conceived as just the objects in the $V_\alpha(U)$s.[47] And he gave his support to the idea that the axioms of first-order axiomatic set theory "underlie" the concept of set. Each of those ideas is tremendously influential today.

The iterative conception gives the Axiom of Foundation center stage: as Zermelo showed, that axiom ensures precisely that each set is a "set of" sets that occurred at previous "layers" or iterations of the "set of" operation. The axiom guarantees that all sets are iterative sets, and the iterative conception makes the axiom obvious.

The iterative conception of set was not, as we have seen, present in the development of set theory up until at least 1930, at least not so far as Zermelo knew. For otherwise he would have introduced the Axiom of Foundation, which is characteristic of the iterative conception, not merely as a provisional restriction for mathematical convenience but as a provisional further specification of what is meant by "set."

Gödel apparently knew of no precedent for the iterative conception either: in the 1960s, when revising his article, he considered crediting Zermelo with "substantially the same solution of the paradoxes,"[48] citing [Zer30], the 1930 article discussed in §4. In the end, quite rightly, he did not. (See Moore's discussion of Gödel's article [Moo86, p. 167].) It was an extraordinarily bold move on Gödel's part to introduce the iterative conception as fundamental so late in the development of set theory.

Someone will surely object that the iterative conception was already implicitly present in Cantor's definitions of a set as "bound up" or "collected." But, as we have seen, Cantor's theory can be spelled out without Foundation. Cantor never, so far as I know, commented on whether a set can be a member of itself. Given the impredicativity of Cantor's theory, there seems to me to be no reason why an enumeration of elements cannot, after the fact, turn out to be such that its range is one of the elements, in which case a non-well-founded set would be a Cantorian set.[49] Moreover, none of Cantor's successors saw such an idea in his work, at least not until after 1947.

Can there be a mathematical argument for the iterative conception? First of all, there can be no proof that the Axiom of Foundation is either true or false. Von Neumann's proof that Foundation cannot lead to contradictions (mentioned in §4) shows that in any model[50] of any of the usual axioms for set theory (without Foundation) the well-founded sets form a model of those axioms plus Foundation. Thus, even when we allow non-well-founded sets it will remain possible to consistently add Foundation by restricting our attention to the well-founded sets. On the other side, results concerning Anti-Foundation mentioned in §4 in fact establish that for any given model[51] of any of the usual axioms plus Foundation, there is a model of Anti-Foundation extending it and such that the given model is the well-founded part of the ex-

[47] The term *iterative* was used earlier by Bernays in connection with set theory in a passage that is in some ways very similar to the one just quoted [Ber35b, p. 260]. Gödel briefly mentioned the iterative conception in print in 1944 [Göd44, p. 462]. The mathematical background for the iterative conception was, as we saw in §4, developed in [Zer30]. But the idea that it is constitutive of what sets are that each one is in some $V_\alpha(U)$, which I take to be the essential component of the iterative conception, was *not* in [Zer30], as we have seen.

[48] No solution of the paradoxes was in fact proposed by Gödel. What was proposed instead is that the paradoxes are not relevant to set theory.

[49] A collection whose only member is itself provides a cheap example: the function that takes 0 to that object witnesses that it is a set.

[50] The result even applies to class models, which will be introduced in §VII.4.

[51] Once again, the result even applies to class models.

tension. Thus, even if we exclude non-well-founded sets it will remain possible to consistently add Anti-Foundation by embedding the well-founded sets into a model of it. The combined effect of these facts is to show that, so far as our present knowledge is concerned, non-well-founded sets are perfectly good mathematical objects and that we should not expect to settle the question whether or not there are non-well-founded sets by proving or disproving Foundation on the basis of some new independently supported axioms.

The only other way to settle the question whether Foundation is true on internal mathematical grounds is in terms of its usefulness, but, as the historical development of set theory makes clear, Foundation has no mathematical use.[52] As Azriel Levy noted in a work that adopted the iterative conception [FBHL73, p. 87], each of the other axioms "was taken up because of its essential role in developing set theory and mathematics in general; if any single axiom were left out we would have to give up some important fields of set theory and mathematics . . . The case of the axiom of foundation is, however, different; its omission will not incapacitate any field of mathematics."[53] It must be granted, however, that it is equally true that the inclusion of Foundation will not incapacitate any field of mathematics: Every structure is isomorphic to a well-founded one, so that when one works only up to isomorphism, as is usual in mathematics, there is no loss in excluding the non-well-founded sets. That provides an excellent justification for Zermelo's policy of adopting Foundation as a simplifying assumption, but it provides only the weakest kind of support for the iterative conception as a conception. The iterative conception entails that it is part of the very idea of what a set is that all sets are well founded. The fact that for many purposes we can live without the non-well-founded sets hardly shows that the very idea of a non-well-founded set is incoherent. (Indeed, it *is* coherent as the arguments mentioned above based on Anti-Foundation show. That is already enough to show in a certain weak sense that the iterative conception is false.)

[52] It has sometimes been claimed that Foundation is needed to obtain an adequate theory of cardinality. Levy's theorem concerning the definability of cardinal numbers is cited [Lev69]. But what Levy showed is that either Foundation or Choice is required to obtain an adequate theory of cardinality. Since for our purposes in this work Choice is always assumed, Foundation is not needed for an adequate theory of cardinality. The most common theory of cardinality is in fact the one we rely on here, the one based on Choice, not Foundation. See, for example, [Vau85] for an elementary treatment.

[53] Levy parenthetically expressed doubts about the necessity of the Axiom of Extensionality.

Whatever defense the iterative conception has must be philosophical. To the extent that the conception integrates and provides a motivation for adopting the axioms of set theory as true of all sets, it is worth adopting. If it showed how to integrate the axioms into a coherent picture of what sets are, the loss of generality occasioned by restricting our attention to well-founded sets might be worth it.

Because the critical issue is the extent to which the iterative conception integrates the theory of sets, I have characterized the iterative conception very narrowly. Many advocates of the iterative conception include in it not only the idea that the sets are built up by iteration of the power-set operation, but also some aspects of what has here been called the combinatorial conception. (As Maddy kindly pointed out to me, Gödel may have been one of those who advocated such a combined picture—that may be the force of "underlying" in the opening sentence of the second paragraph of the passage quoted earlier.) But advocates of the kind of mixed view just described cannot claim that it provides an integrated theory of sets without telling a story about how the two aspects of their view come naturally together—a mere conjunction is not enough. No such story has been provided. I therefore adopt a narrow characterization of the iterative conception, which makes it possible to carefully analyze how that conception fits in with ideas drawn from the combinatorial conception.

So let us see how much the iterative conception helps us to integrate and motivate the axioms of set theory.[54] "Sets of" objects are certainly sets in some antecedent sense. We must therefore adopt the Axiom of Extensionality, not motivated by the iterative conception but prior to it. We are instructed by the iterative conception to take all sets of sets obtained by iteration, but the instruction "all sets" is of no help without some prior understanding of what sets there may be, an understanding that requires, at least, the Axiom of Separation. The Axiom of Separation does not follow from the iterative conception either. Like Extensionality, it must be part of a prior concept of set.

The Axiom of Choice also depends on the prior concept of set. If Choice is true of sets, then it will, on the iterative conception, be true of sets obtained by iterating the "set of" operation: Let S be a nonempty iterative set of pairwise disjoint iterative sets. If the Axiom of Choice is true, then there will be a set T that contains exactly one member of each member of S. Since each member

[54] Many expositions of the iterative conception have appeared. For critical discussions of some of the most important ones, see [Hal84, especially pp. 214–223] and [Par77].

of a member of an iterative set is itself an iterative set, and since all of the members of members of S appeared at some stage before S, the set T will be an iterative set that appears by the stage at which S does. Thus, if the Axiom of Choice is true of sets, it will also be true of iterative sets. But that is of no help in determining whether or not the Axiom of Choice *is* true. To the extent that the iterative conception is an autonomous conception, it fails to help in settling whether or not Choice holds. Choice must be taken as an additional assumption, as some advocates of the iterative conception have noted.[55]

The iterative conception does not tell us how far to iterate (see Gödel's note *b*), and so we must also start with an Axiom of Infinity. In addition, for the same reason, the iterative conception presupposes the notion of "transfinite iteration." In effect, the ordinal numbers are supposed given in advance.[56] One of the symptoms of the need to start with whatever ordinal numbers are used is that the Axiom of Replacement, which, as we have seen, is intimately associated with the ordinal numbers, is not a consequence of the iterative conception.

The remaining axioms do follow from the iterative conception:

Pairing. If s and t are two iterative sets, then the set of s and t is an iterative set formed at the first stage after both s and t are.

Union. If S is an iterative set, then its members are all iterative sets formed before S and their members are all iterative sets formed before they

were, and hence before S. We can therefore form the set of members of members of S at the same stage we form S.

Power Set. Any subset of an iterative set S will be formed by the stage at which S is. We can therefore form the set of all of them at a stage after S is formed.

The reader should by now be uncomfortable with my talk of "forming" some sets "before" or "after" others. It is crucial to maintaining the full impredicative forms of Separation and Replacement that sets be construed realistically, not as being constructed by us, as Zermelo essentially argued in 1908 (§1). So whatever notion of priority is being invoked by the iterative conception cannot be temporal, and whatever notion of formation is being invoked cannot be construction. Moreover, as Charles Parsons pointed out [Par77, p. 507], the temporal metaphor of a mind collecting objects already constructed breaks down for nondenumerable iterations and collections: "It is hard to see what the conception of an idealized mind is that would fit here; it would differ not only from finite minds but also from the divine mind as conceived in philosophical theology, for the latter is thought of either as in time, . . . or its eternity is interpreted as complete liberation from succession." Even a Cantorian appeal to God's powers proves to be inadequate here!

Without the temporal notion of constructing some sets after others the iterative conception loses much of its appeal. Parsons has suggested that a modal interpretation of the iterative conception be given to avoid the reliance on time: a multiplicity of actual sets is a possible set ([Par77, pp. 509, 515], [Par83c]). That is a fascinating suggestion well deserving of further exploration. It is not yet, however, clear how to interpret the necessary notion of possibility without relying on metaphors of time and construction.

At least at present the advantages of the iterative conception do not suffice to justify adopting it: It does not provide a conception that unifies the axioms of set theory. It is based on a picture of construction that does not seem to have a coherent interpretation. And, most damning of all, even though the iterative conception has been widely embraced in recent times, it has had very little impact on what theorems can be proved—no essential mathematical use has been found for Foundation.[57]

[55] Hao Wang [Wan74] called his concept of set the iterative conception. In my terminology, he—like others—combined a combinatorial conception of set, the iterative conception, and some elements of limitation of size to motivate the axioms of set theory, including Choice and Separation. But Choice and Separation follow from the combinatorial aspect of his conception, not the, in my terms, iterative one. Wang described intuitions behind the axioms of set theory in a way that does capture, I believe, a lot of the picture that goes with the present widespread acceptance of those axioms, but the iterative aspect is not the crucial one.

[56] George Boolos [Boo71] and Dana Scott [Sco74] have given variants of the iterative conception that avoid the need for that. Boolos assumed in his formal theory that one can prove things about the stages by induction instead of explicitly assuming that the stages are constructed in a sequence indexed by ordinal numbers. But the assumption about induction is motivated, as Boolos clearly stated, by a "rough description" that does involve the ordinals. Scott employed a reflection principle. While his axiomatization of set theory is interesting in its own right, the use of reflection principles is not a part of the notion of "set of," and so his axiomatization is not relevant to an evaluation of the basic iterative conception.

[57] Given the widespread employment of Foundation today it remains possible that some use will yet be found. In that case, I would reverse my position. Caution therefore dictates that I not make my other arguments depend on a denial of the iterative conception, and indeed none of them do depend in any way on such a denial.

There is a final advantage of the iterative conception that is best brought out by contrasting it with the limitation-of-size conception. The limitation-of-size conception, recall, has it that if a collection is the same size as a set, then it is a set. Just as the Axiom of Foundation is characteristic of the iterative conception, the Axiom of Replacement is characteristic of limitation of size. But note that limitation of size tells which among the collections are sets—those that are small enough. Unless one just takes Replacement to express the limitation-of-size idea, limitation of size suggests, on pain of vacuity, that there are some collections that are not sets. Thus, for example, it is immediate from von Neumann's Limitation of Size Axiom that the universe is a class that is not a set. Though "set of" requires an antecedent notion of set, it does not require that there turn out to be collections that are not sets. If, like Gödel, one wants a domain of all sets, then one based on limitation of size will plausibly involve collections that are not sets,[58] while one based on the iterative conception could be construed as a domain of all collections as well as all sets. That is, the iterative conception leaves open the possibility of claiming that there is no collection of all sets and that every collection is a set.

If one is willing to give up on a single domain of all sets, as Zermelo was, then limitation of size is compatible with the claim that all collections are sets. It is just not compatible with the claim that all collections are sets in a single domain. Nonetheless, we can see immediately that limitation of size will not serve as an overall guiding principle for our set theory any more than the iterative conception does: limitation of size does not justify the Power Set or Union Axioms. It is not clear that it justifies Choice either: Given a set S of disjoint nonempty sets, a collection T consisting of exactly one object from each of them will be the same size as S and will therefore be a set—if it exists. The truth of Choice depends on our antecedent theory of collections, which is not given by limitation of size. There is, however, as we have seen

[58] Given suitable background assumptions, one *can* formulate limitation of size without classes. One version, suggested to me by Vann McGee, is that given a well-ordering of the universe, limitation of size can be formulated as follows: If there is an x such that every y such that $\phi(y)$ is less than x, then there is a set of all ys such that $\phi(y)$. (Naturally, I have in mind a schema in which formulas with no free occurrences of x may be substituted for ϕ.) I do not know of a proposal along such lines that seems sufficiently natural to merit serious consideration, and so in the text I have pretty much ignored the possibility of formulating limitation of size without classes.

in §3, an extension of limitation of size from which Choice does follow: von Neumann's Limitation of Size Axiom.

There is another limitation theory besides limitation of size: Fraenkel's theory of limitation of comprehensiveness, suggested to him by Zermelo's Separation Axiom. Since it played no role in the development of the axioms, I have not discussed it before. But Fraenkel used various versions of it to justify the axioms from 1924 through at least 1958. See [Hal84, pp. 200–202]. The basic idea is that, starting with given sets, we only form sets from them that are somehow connected to them (how is never exactly clear), so that the sets formed are not absolutely comprehensive like the sets of the paradoxes. Here is a version—my own, not Fraenkel's: One way to guarantee that a collection is not too comprehensive is to require that its members already be bounded by a set. That idea suggests the following axiom: every subcollection of a set is a set. Given a reasonable notion of collection that yields the Separation Axiom. But it doesn't get us much else. However, it seems a reasonable extension of the idea to allow that a collection is not too comprehensive, and hence forms a set, if the members of its members are bounded by a set. That suggests the axiom: a collection is a set if its union is included in a set. That, in combination with Separation,. which we have just seen follows from limitation of comprehensiveness, yields Power Set. It should be fairly clear from the above that limitation of comprehensiveness shares some similarities with the iterative conception. Gödel's note c in the quote above is, as Hallett suggested [Hal84, p. 202], reminiscent of the limitation-of-comprehensiveness.

Just as in the case of limitation of size, assuming the converse of the limitation of comprehensiveness seems like a reasonable extension: the sets are exactly the collections that are not too comprehensive. The converse of the limitation-of-comprehensiveness principle that yielded Power Set yields Union.

We can combine the two limitation theories to obtain an axiomatization of set theory that is as well motivated as any. Since both theories start with collections and delimit the sets among them, the theory will—like von Neumann's in the usual presentation—be a theory of classes with axioms to tell which classes are sets.

As in the modified von Neumann theory, a set, by definition, is a class that is a member of a class. The first axiom is Extensionality for classes.

AXIOM 5.1. *Classes with the same members are equal.*

We shall also need to have axioms to ensure that there are enough classes.

Von Neumann's axioms on classes would do, as would those of Bernays [Ber76, p. 5] or Gödel's based on them [Göd40, p. 37], but those axioms all assume the existence of ordered pairs of sets that are sets, and so I prefer the following axiom schema, in which $\phi(x)$ is any definite formula:[59]

AXIOM 5.2. *There is a class with members exactly the sets x such that* $\phi(x)$.

I used the vague term *definite* because there are several ways of spelling it out. There is the first-order version of the axiom in which a definite ϕ is simply a first-order formula in the language of the theory with quantification only over sets. That version is the "class theorem" of Bernays [Ber76, pp. 12–13]. There is Zermelo's second-order version, in which the domain of a model is a set in a higher model. There, the definite formulas will also include ones of the form $x \in S$, where S is an arbitrary subset (in the sense of a higher model) of the domain. There will be a first-order version that allows quantification over classes, a first-order version for any expanded language, and there is at least one other version as well, to be discussed in §VII.4. In any event, the axiom will serve to guarantee that the classes are closed under simple operations. In particular, if S is a class, then so is its union, the class of members of members of S. We need an Axiom of Infinity. I like this one:

AXIOM 5.3. *There is a nonempty set on which the membership relation is a discrete linear order with no last element.*

Next, we have von Neumann's Limitation of Size Axiom 3.1. Last of all, we have the following:

AXIOM 5.4 (LIMITATION OF COMPREHENSIVENESS). *A class is a set if and only if its union is a set.*

Technical Remark. The above axiom system (in its first-order version) is equivalent to von Neumann's, reformulated as usual, as is easily seen: Limitation of Comprehensiveness yields Union and Power Set, and conversely. The version of the axiom system along Zermelo's lines has as models those Zermelo noted were models of von Neumann's system: domains of the form V_κ, where κ is strongly inaccessible and the sets are exactly the members of $\bigcup_{\beta<\kappa} V_\beta$.

The "only if" of Limitation of Comprehensiveness is redundant. That direction is the Union Axiom; the remaining direction is what gave Power Set. But Levy showed that Limitation of Size plus Power Set yields Union [Lev68].

First order or second order? One intended domain or many? Are all collections sets? Do the iterative conception, limitation of size, and limitation of comprehensiveness fit into a coherent conception of sets? The last question is the heir of Cantor's quandary, How does Power Set fit into a combinatorial conception of sets? Our understanding of the foundations of set theory is not much better than d'Alembert's understanding of the foundations of analysis was in the latter half of the eighteenth century.[60]

[59] I have excluded urelements to simplify the axioms. A fully general version should allow them. As usual, there is no essential problem in doing so, but some minor increases in complexity result. I leave the needed changes to the reader.

[60] Fraenkel expressed a similar sentiment in 1927 [Fra27, p. 61]. The situation has not changed all that much since.