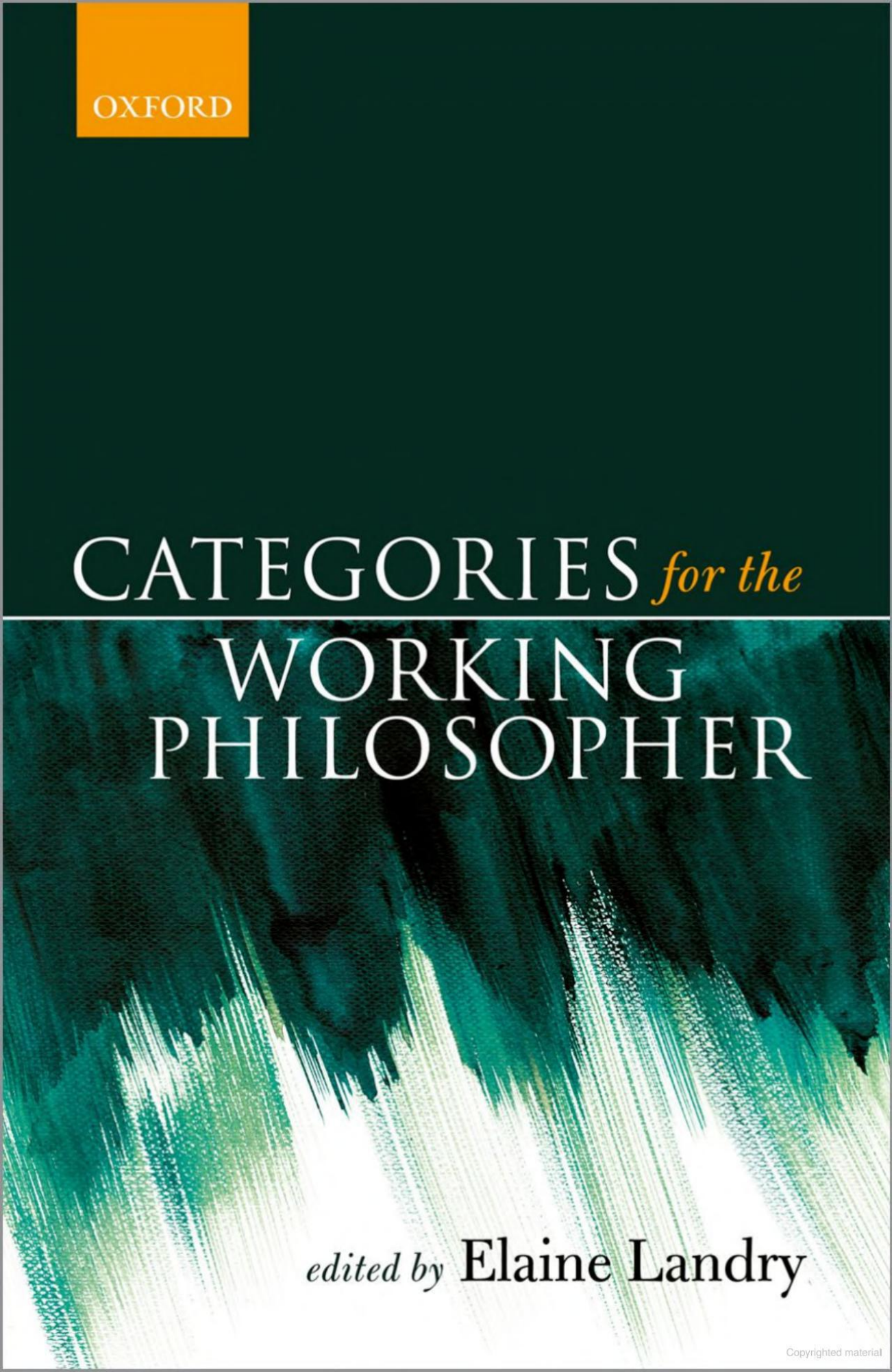


Categories for the Working Philosopher

EDITED BY
Elaine Landry

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 2017

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2017940285

ISBN 978–0–19–874899–1

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

Notes on Contributors

Samson Abramsky is Christopher Strachey Professor of Computing at the Department of Computer Science, University of Oxford. His speciality areas are concurrency, domain theory, lambda calculus, semantics of programming languages, abstract interpretation, and program analysis.

Steve Awodey is a professor at the Department of Philosophy, Carnegie Mellon University. His specialty areas are category theory, logic, philosophy of mathematics, and early analytic philosophy.

John L. Bell is a professor in the Department of Philosophy, University of Western Ontario. His specialty areas are mathematical logic, philosophy of mathematics, set theory, boolean algebras, lattice theory, and category theory.

J. R. B. Cockett is a professor in the Department of Computer Science, University of Calgary. His specialty areas are distributive categories, restriction categories, linearly distributive categories, differential categories, categorical proof theory, semantics of computation, semantics of concurrency, categorical programming, and quantum programming.

Bob Coecke is Professor of Quantum Foundations, Logics, and Structures at the Department of Computer Science, University of Oxford. His specialty areas are foundations of physics, logic, order and category theory, and compositional distributional models of natural language meaning.

David Corfield is a senior lecturer in philosophy at the University of Kent. His specialty areas are philosophy of science and philosophy of mathematics.

Andrée Ehresmann is Professeur Emérite at the Department of Mathematics, Université de Picardie Jules Verne. His specialty areas are analysis, category theory, self-organized multiscale systems with applications to biology and cognition.

Michael Ernst is a graduate student in the Department of Logic and Philosophy of Science, University of California, Irvine. His specialty area is philosophy of mathematics.

Hans Halvorson is a professor in the Department of Philosophy, Princeton University. His specialty areas are category theory, logic, philosophy of science, and philosophy of physics, science, and religion.

Kohei Kishida is a post-doctoral researcher in the Department of Computer Science, University of Oxford. His specialty areas are category theory, philosophical logic, and modal logic.

Aleks Kissinger is a post-doctoral research assistant in the Department of Computer Science, University of Oxford. His specialty areas are category theory, quantum information, graphical calculi, and graph rewriting.

Joachim (Jim) Lambek was an emeritus professor, Department of Mathematics and Statistics, McGill University. His specialty areas were categorical logic and mathematical linguistics.

Elaine Landry is a professor at the Department of Philosophy, University of California, Davis. Her specialty areas are philosophy of mathematics, philosophy of science, analytic philosophy, philosophy of language, and logic.

Jean-Pierre Marquis is Professeur titulaire, Department of Philosophy, Université de Montréal. His specialty areas are logic, philosophy of mathematics, philosophy of science, foundations of mathematics, and epistemology.

Colin McLarty is Truman P. Handy Professor of Philosophy & Professor of Mathematics, Department of Philosophy, Case Western Reserve University. His specialty areas are logic, philosophy of logic, philosophy of mathematics, philosophy of science, and contemporary French philosophy.

R. A. G. Seely is an adjunct professor at the Department of Mathematics, McGill University. His specialty areas are categorical logic, linear logic, computational logic, and proof theory.

Michael Shulman is an assistant professor at the Department of Mathematics and Computer Science, University of San Diego. His specialty areas are category theory and algebraic topology.

David I. Spivak is a research scientist at the Department of Mathematics, Massachusetts Institute of Technology. His specialty areas are category theory, categorical informatics, and derived manifolds.

Dimitris Tsementzis is with the Department of Philosophy, Princeton University. His specialty areas are logic, foundations of mathematics, and category theory.

James Owen Weatherall is an assistant professor in the Department of Logic and Philosophy of Science, University of California, Irvine. His specialty areas are philosophy of physics, philosophy of science, and mathematical physics.

1

The Roles of Set Theories in Mathematics

Colin McLarty

> This examination consists in specifying topics within mathematics for which the appropriate branches of logical foundations [l.f.] do or do not contribute to effective knowledge.
>
> Correspondingly, the demand, accepted (uncritically) in the early days of l.f., for foundations of all mathematics, by logical means to boot, is replaced below by a question: In which areas, if any, of mathematics do such foundations contribute to effective knowledge?
>
> (Kreisel, 1987, 19)

Like Kreisel (1987), we here do not argue about logical foundations for all mathematics. We look at how specific set theories in fact advance mathematics. However, Georg Kreisel's great concern was *effective* knowledge in the logician's sense of finding explicit numerical solutions or at least explicit numerical bounds on solutions to arithmetic problems. Here I use "effective" in the colloquial sense of widely successful in producing a desired or intended result, whether or not it is specifically a numerical solution to an arithmetic problem.

Let me explain because philosophers sometimes miss this topic. I do not ask here what kind of foundations are necessary in principle, nor what all ideas of sets have been used for something at some time, nor what might lead to progress in the future. Those fine questions are not the topic here. This paper addresses set theories in widespread, currently productive use. Excursuses 1.2.1 and 1.2.2 on the Continuum Hypothesis and Grothendieck Universes discuss two interesting gray areas where it matters just how widespread and productive you want it to be

1.1 Overview

No one can be surprised that the role of set theory in mathematics varies with the kind of mathematics. Topologists and analysts face set-theoretic issues, notably the

Continuum Hypothesis, which number theorist do not. Some people will be surprised, and may even object, on hearing different set theories are typical in different parts of mathematics. Linnebo and Pettigrew (2011) apparently make a counterclaim:

> Many textbooks that introduce elementary areas of mathematics, such as algebra, analysis, and number theory, include an elementary section surveying the elements of set theory, and this is explicitly orthodox set theory. (249)

But they do not say what they mean by "orthodox". It is true that textbook set theory is rarely intuitionistic, predicative, modal, or non-classical in other ways familiar to philosophers. And in logic texts it is almost always Zermelo Fraenkel set theory, which philosophers tend to take as orthodox. But texts outside of logic rarely come close to Zermelo–Fraenkel.

Section 1.2 gives a few reasons why "orthodox set theory" does not per se mean Zermelo–Fraenkel set theory. Section 1.3 looks at set theory in two standard first year graduate textbooks: James Munkres' *Topology* (2000) and Serge Lang's *Algebra* (2005).[1] It also looks at the recent article by mathematician Tom Leinster (2014). All support the claim that mathematicians know and use the concepts and axioms of the Elementary Theory of the Category of Sets (ETCS), often without knowing or caring that they are the ETCS axioms.

Section 1.4 uses a concept of "mathematical gauge invariance" to show why the category of sets described in ETCS is a closer fit to the practical needs of most mathematicians than is the cumulative hierarchy of sets described in ZFC. For example, philosophers of science may suspect that nothing in mathematical practice depends on solving the multiple reduction problem of whether the number 2 *is* the ZFC set $\{\{\phi\}\}$ or the set $\{\phi, \{\phi\}\}$, or some other set. Few mathematicians have ever heard of this alleged problem.

Mathematicians do constantly meet a similar problem: take the tangent bundle $T(M)$ of a manifold M as a geometric example. Spivak (1999, ch. 3) gives three quite different geometric constructions starting from M. The three give naturally isomorphic results, any one of which will be used as "the" tangent bundle $T(M)$ of M for some purposes. Let me stress: the results are not just different sets when formalized in ZFC. They rely on different aspects of geometry. So the difference between them matters in geometry. Geometers daily rely on nontrivial theorems showing the results are isomorphic, which is why Spivak spends a chapter on the definitions and the proofs.

For this and many other reasons geometers have developed agile, rigorous techniques for handling spaces that are only defined up to isomorphism. Most fields of

[1] Some may object that I only consider two cases. They are influential books, and McLarty (2008a) and (2012) discuss several more. But really that objection gets things backwards because no matter how many books I cite not using Zermelo–Fraenkel theory with choice (hereafter ZFC) one could suspect I left out scores more that do. Rather, philosophers who believe many mathematics texts use ZFC should specify at least a few and show how those use ZFC any more than Munkres (2000) does.

mathematics rely on such techniques . These are not the techniques of philosophical structuralism! They are categorical and functorial techniques. ETCS was created from these same techniques.

Section 1.5 describes the positive uses of the extra structure of *an iterated hierarchy* which is assumed for ZFC sets, and compares it to the yet further structure of *constructibility*. Constructibility in this sense was introduced by Gödel (1939, 577) as "a new axiom [which] seems to give a natural completion of the axioms of set theory" and which implies the Continuum Hypothesis. Today constructibility is assumed in most uses of set theories adapted to arithmetic. But it is incompatible with most kinds of large cardinals, so ZFC set theorists treat it as a technical tool and not a property of all sets, as did Gödel himself as he soon rejected the view he had taken in 1938.

1.2 Why ZFC is not Synonymous with "Set Theory"

Zermelo–Fraenkel set theory with choice has indeed been orthodox in set-theoretic research especially since Cohen (1966) used it for the method of *forcing*. But other parts of logic use other set theories. For example, Kreisel's work on effective knowledge led to an array of set theories at the far extreme from ZFC, and closely related to arithmetic, as expounded by Hájek and Pudlák (1993, ch. 1) and Simpson (2010, ch. 1). These theories are provably too weak for some standard mathematics but that is exactly what adapts them to elucidating effective knowledge in the logical sense, that is in Kreisel's sense.

Outside of research logic most mathematicians succeed in the practice of mathematics without ever seeing the ZFC axioms or the set theories close to arithmetic. Few mathematicians could state axioms for any set theory, and more than a few insist this is as it should be. When Alexander Grothendieck began creating the now standard tools of algebraic geometry he used a large cardinal axiom added to a set theory similar to ZFC, and he commissioned a 40-page exposition of it by N. Bourbaki (Artin et al., 1972, 185ff.).[2] A number of number theorists have deplored, not the new axiom itself which is modest by set theorists' standards for large cardinals, but the very idea of mixing axiomatic set theory with number theory. Excursus 1.2.2 returns to this.

Could it be that most mathematicians use ZFC the way most people drive a car without knowing how the engine works? That is, are most mathematicians content to sketch how proofs of their theorems should go, while letting others, trained in logic, worry about actual proofs? Certainly mathematicians often use results in mathematics that they have never seen fully proved.

The extent of this varies with the field. When one algebraist said he would never cite a theorem in a paper unless he knew the proof, my teacher Charles Wells responded, "We can do that in abstract algebra. Differential geometers have to use more theorems

[2] Bourbaki is the pseudonym of a group of mathematicians. Pierre Cartier, who was in the group at the time, says Pierre Samuel wrote this appendix (conversation February 19, 2015).

than they have personally checked". But on one hand neither of those algebraists took proof to mean a proof in ZFC. And more to our point the results a mathematician will cite in research without learning the proofs are generally not basic facts used throughout the work. They are more special theorems. Mathematicians normally can prove the facts they use daily. And I will maintain this includes the facts of set theory that they actually use. An analyst, algebraist, or topologist sometimes mentions an independence result in set theory without having worked through its proof. But they rarely *use* independence results.

A philosopher of mathematics might take a hard stand, saying most mathematicians do not and cannot justify their theorems: because justification requires proofs in ZFC which few mathematicians know. The philosopher could say nonetheless the theorems are justifiable because philosophers and logicians can recast the proofs in ZFC. I cannot argue against such epistemology here. But I prefer the hard-won insight of Russell (1924, reprint 326) that far from foundations justifying mathematics, foundations themselves must be "believed chiefly because of their consequences" in mathematics.

Alternatively, could it be like speaking in prose without knowing it is called "prose"? Perhaps mathematicians generally know the ZFC axioms without knowing they are the ZFC axioms? That is more in line with Russell's mature thinking. And indeed most mathematics texts say a lot about sets without calling anything an axiom of set theory. But the things they say are not distinctive of ZFC. Far from tacitly knowing the ZFC Axiom of Foundation, or the Axiom Schemes of Separation and Replacement, most mathematicians are unfamiliar even with the notions of *global membership* and *first-order axiom scheme*. Those notions are basic to stating these axioms which are in turn basic to ZFC. We will see that mathematicians do know the concepts and axioms of the *Elementary Theory of the Category of Sets*, which, after all, Lawvere (1964) took from the practice of mathematics around him.

1.2.1 The Continuum Hypothesis

Among Cantor's first decisive achievements was his proof that the real numbers $\mathbb{R}$ are *uncountable*: they cannot be put into one simply infinite sequence $r_0, r_1, \ldots$. He then conjectured that $\mathbb{R}$ has the smallest uncountable cardinality: every infinite subset $S \subseteq \mathbb{R}$ either is countable or can be put in bijection $S \cong \mathbb{R}$ with the set of all reals. This is Cantor's Continuum Hypothesis (CH). It remains a leading topic of set theory and arguably the leading topic ever since Cantor, because no generally accepted set theory can either prove it or refute it. It is not just that no one knows a proof or refutation in any accepted set theory. Masses of research descended from Gödel (1939) and Cohen (1966) prove all generally accepted set theories are consistent with both the truth and the falsity of CH. Nearly all of this set theoretic work is conducted in terms of ZFC and extensions of ZFC.

This is obviously relevant to analysis, and while it might be surprising how rarely it matters, it does come up sometimes. For example, it matters exactly once in Munkres'

Topology. An exercise, mentions a question about *box topologies* whose answer is not known, though it is known that "the answer is affirmative if one assumes the Continuum Hypothesis" (Munkres 2000, 205). He cites a proof by Mary Ellen Rudin (1972), which is also inexplicit about its set theory. Rudin cites Bourbaki and also Kelley (1955), who each give their own set theories (Kelley's is stronger than ZFC). But she says nothing about either one's set theory. Her topological argument is insensitive to any difference among ZFC, ETCS, and those others.

The long and short of it for Munkres was that this question on box topologies could only be settled by making a controversial assumption, the CH. He felt topology students should know this happens sometimes.

1.2.2 Grothendieck universes

Grothendieck found penetrating new ways to calculate with small (often finite) structures, by organizing those small structures into very large categories and functors, e.g., *Abelian categories* and *derived functors*. He was not very interested in set theory, yet thought it was worth getting right, and he saw that his large categories and functors were far too big for ZFC to prove they exist. He did not only use collections such as the class of all sets, which is too big to be a set. He used collections of those too-large collections, and larger-yet collections of those, and more. McLarty (2010) gives details for logicians.

Grothendieck cared enough to give a rigorous set-theoretic foundation. But he did not linger on minimizing the foundation in any way. He gave a quick solution using *Grothendieck universes*, which are sets so large that ZFC does not prove they exist.

The impact of this set theory on mainstream mathematics is attenuated by a mix of factors:

1. While Grothendieck's cohomology is entirely standard in research number theory and geometry, it remains a rather advanced specialty.
2. As Grothendieck intended, the large structure tools are not the focus of attention but merely a framework organizing calculations.
3. As Grothendieck knew, those calculations can be done without the large structures, at the loss of conceptual unity and general theorems.
4. Texts such as Milne (1980), Freitag and Kiehl (1988), and Tamme (1994) use large structure theorems informally without discussing foundations.
5. Others, like (Fantechi et al., 2005, 10) and Lurie (2009, 50f.), invoke Grothendieck universes only to dismiss a technical problem.
6. The large structures can be founded on a conservative extension of ETCS, far weaker than ZFC, let alone a Grothendieck universe (McLarty, 2011).

In sum, this has not made axiomatic set theory a standard topic in geometry or even in specifically Grothendieck-inspired geometry.

1.3 Sets for the Working Mathematician

André Weil, whose research in number theory and geometry put him among the leading mathematicians of his time even apart from his role as the initiating member of Bourbaki, launched the meme "X for the working mathematician" with a talk titled "Foundations of Mathematics for the Working Mathematician", delivered to the Association for Symbolic Logic (1949). He aimed to bring mathematical practice closer to logical principles. However much people complain about Bourbaki's abstractness, or neglect of geometry or of advanced logic, it remains that Bourbaki was a leader in making mathematical writing radically more accessible. And they did this using set theory. Mathematics textbooks and published research became more uniform in style, and more rigorous, than before World War II. The spread of explicit set theory was a great part of this, going hand in hand with the demand for readily readable textbooks and more uniform terminology in the burgeoning new fields of mathematics (McLarty, 2008b). This has become the norm in graduate mathematics teaching worldwide, because experience showed it made the subject easier to learn.

It is valuable to match a modern introduction to differential geometry with classical readings from Riemann, for example, as Spivak (1999) does. But few people today could read Riemann without the help of modern texts. Few ever did succeed at reading Riemann in the nineteenth century, or in the first half of the twentieth century. It is easier today because of the set-theorization of mathematics associated with Bourbaki among others. Nostalgia for the good old days of easy, intuitive mathematics is misplaced.

Bourbaki (1958) created their own set theory similar to ZFC. But their volumes on various fields of mathematics rarely referred to it and few mathematicians, or even logicians, ever learned it.

To see how the set-theorization actually happened in widespread practice let us look at set theory in two currently influential graduate textbooks. Neither of them gives precise axioms. Rather they discuss more or less basic facts about sets which they go on to use. We focus on two questions bearing on the distinction of ZFC from ETCS set theory:[3]

1. How does the book handle the elements of sets?
2. Does the book define functions as a kind of set?

Munkres (2000) opens with seventy leisurely pages:

[3] As a more technical issue: every textbook I looked at freely forms sets of sets without discussing the Axiom Scheme of Replacement. One explanation would be that these avowedly naive treatments accept unlimited comprehension though that is actually inconsistent. Or one might notice these apparent uses of replacement can be reduced to the more innocuous Bounded Separation Axiom Scheme, which in turn can be stated as a single axiom. There is usually a suitable ambient superset ready at hand. The point here, though, is that none of these textbooks is sufficiently explicit about logic even to state Replacement or Separation axiom schemes, or to explain the difference between an axiom and an axiom scheme. Those schemes are not part of ETCS, though they are available there as add-ons if wanted.

I begin with a fairly thorough chapter on set theory and logic. It starts at an elementary level and works up to a level that might be described as "semisophisticated". It treats those topics (and only those) that will be needed later in the book. (2000, xi)

In order to "introduce the ideas of set theory", Munkres says, "Commonly we shall use capital letters $A, B, \ldots$ to denote sets, and lowercase letters $a, b, \ldots$ to denote the objects or elements belonging to these sets" (2000, 4). Of course in ZFC everything is a set, including that the elements of sets are sets. Munkres hardly denies that all objects are sets. But, just as he says he will do, he commonly writes as if elements are of a different type than sets, which they indeed are in ETCS.

Munkres comes close to ZFC practice on cartesian products. He says $A \times B$ is the set of all ordered pairs $\langle a, b\rangle$ with $a \in A$ and $b \in B$. He treats it as optional to say that $\langle a, b\rangle$ is a set itself. He notes it can be defined as the set $\{\{a\}, \{a, b\}\}$ and he shows that this set uniquely determines a and b. But he concludes, "it is fair to say that most mathematicians think of an ordered pair as a primitive concept rather than thinking of it as a collection of sets" (2000, 13).

Munkres says mathematicians think of a function $f : A \to B$ as a rule assigning values $f(a) \in B$ to arguments $a \in A$, but they also need a more precise definition (2000, 15). For this he follows the ZFC practice of defining a function $f : A \to B$ as a set. Namely, f is a subset of the product $f \subseteq A \times B$ with the property that for each $a \in A$ there exists a unique $b \in B$ with $\langle a, b\rangle \in f$.

He comes to "what we might call the mathematical foundations for our study—the integers and the real number system" (2000, 36). He notes there are two approaches to this. One is to construct these sets by building them up from the empty set. That is the standard approach of ZFC, though Munkres does not say so. He does say that approach "takes a good deal of time and effort and is of greater logical than mathematical interest" (2000, 36). He will rather "assume a set of axioms for the real numbers and work from these axioms" without proving from any more basic axioms that any such set exists.

He never entertains the question of what set is 2. For him 2 is an integer, the sum of 1 and 1, it is an element of the set $\mathbb{R}$ of real numbers. He never affirms or denies that 2 is itself a set. We have just seen he declines to give any set theoretic specification of the elements of the set $\mathbb{R}$. Nor does he ever specify what sets are the elements of the sets $\mathbb{Q}, \mathbb{C}$ of rational and complex numbers, respectively. This has a practical advantage as it allows Munkres to assume actual subset inclusions from the natural numbers through the integers, the rational, and the real numbers, up to the complex numbers:

$$\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}.$$

Most ZFC definitions make these not actually subset inclusions. We return to this at the start of Section 1.4.

This is the second edition of a book Munkres first published in 1975, when he had probably not heard of ETCS. There is no evidence he ever heard of any axiomatic set

theory but ZFC. On the other hand, in this book he never actually gives the central concepts or axioms of ZFC. When he does bring up the occasional idea from ZFC he always says mathematicians do not usually think that way.

Lang (2005) is greatly expanded from the first edition (Lang, 1965). The treatment of sets did not change significantly. Lang has none of Munkres's switching between how he believes mathematicians think and what he believes rigor requires. He simply gives the principles he uses in the rest of his book. It is possible that by 1965 Lang had heard of ETCS in conversation with Peter Freyd but also possible he had not. And his book had to be well underway before Lawvere had shown ETCS to anyone at all. In any case I do not claim Lang cared about ETCS. His account simply says nothing that would distinguish between ZFC and ETCS, because nothing in his book depends on those differences.

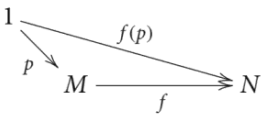
The discussion of sets begins: "We assume that the reader is familiar with sets" (Lang, 2005, xi). Yet Lang proves very simple facts. An appendix proves, for example, that the product of a non-empty finite set with a countably infinite one is countably infinite (878). Without ever saying what a function *is*, indeed without saying functions are sets at all, Lang says what he needs about them: a function $f : A \to B$ may also be called a *mapping*, it has a domain A and a codomain B, and it takes a well-defined value for each element of the domain. His entire explicit discussion of the term is the following:

> If $f : A \to B$ is a mapping of one set into another, we write
>
> $$x \mapsto f(x)$$
>
> to denote the effect of f on an element x of A. (Lang, 1993, ix)

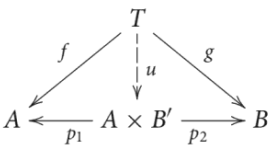
Either ZFC or ETCS would spell this out in just a few steps:

1. ZFC formalizes function evaluation by taking elementhood as primitive and defining an ordered pair as a set by some means such as $\langle x, y\rangle = \{\{x\}, \{x, y\}\}$. Then prove that indeed on this definition $\langle x, y\rangle = \langle w, z\rangle$ implies that $x = w$ and $y = z$. Then define a function f as a suitable set of ordered pairs $f \subseteq A \times B$, and define $f(x) = y$ to mean $\langle x, y\rangle \in f$.
2. Categorical set theory takes function composition as primitive, posits a set 1 such that every set has exactly one function $A \to 1$, and defines elements $x \in A$ as functions $x : 1 \to A$. This implies that 1 has exactly one element, namely the sole function $1 \to 1$. Then for any function $f : A \to B$ and element x of A, define $f(x) \in B$ as the composite $fx : 1 \to B$.

Either way works formally and mathematicians generally learn set theory without ever hearing a formal logical treatment of either one. I will mention that in some parts of geometry a point p of a space M is defined as a function $p : 1 \to M$ from a one point space, so that a function $f : M \to N$ acts on points $p : 1 \to M$ by composition, giving $f(p) : 1 \to N$.



Lang never defines cartesian products $A \times B$ in a ZFC way. Rather, he defines $A \times B$ up to isomorphism in his introduction to "categories and functors" (2005, 53–65). He handles products of sets along with products of groups and other structures by the standard commutative diagram definition (2005, 55). The product is not just a set but a set $A \times B$ plus two functions $p_1 : A \times B \to A$ and $p_2 : A \times B \to B$. And given any set T and functions $f : T \to A$ and $g : T \to B$ there is a unique function u making all the composites in this diagram equal:



$$p_1 u = f \quad \& \quad p_2 u = g.$$

A convenient notation has $u = \langle f, g \rangle$. In that notation, the special case $T = 1$ of the definition says the elements of $A \times B$ are precisely the pairs $\langle x, y \rangle$ with $x \in A$ and $y \in B$.

Again let us be clear about the issue: every mathematician must know cartesian products have this property. It is implicit in all use of products since, say, Descartes, long before the notion of a function or of a cartesian product could be clearly stated. In fact, this property determines the product triple $A \times B, p_1, p_2$ uniquely up to isomorphism, as proved by Lang and by any introduction to category theory. And, in fact, this is the property of products that Lang uses throughout his book. That is why he gives it.

So, can these properties that Lang and Munkres and others actually use be stated precisely in just the terms that Lang and Munkres use? Or must they remain a rough guide which actually requires ZFC for a precise statement? That is exactly what Lawvere asked himself in the years leading up to his Lawvere (1964). For this history, see McLarty (1990).

It is not that Lang, or Munkres, or many other authors chose to use this set theory. Rather this set theory is based on the very techniques that mathematicians use in algebra, topology, and so on, every day.

We need not look at ETCS in all formality. That is done in Lawvere (1964) and many other places. We will just quote the less formal summary given by Leinster (2014, 404):

1. Composition of functions is associative and has identities.
2. There is a set with exactly one element.
3. There is a set with no elements.
4. A function is determined by its effect on elements.
5. Given sets X and Y, one can form their cartesian product $X \times Y$.
6. Given sets X and Y, one can form the set of functions from X to Y.

7. Given $f : X \to Y$ and $y \in Y$, one can form the inverse image $f-1(y)$.
8. The subsets of a set X correspond to the functions from X to $\{0,1\}$.
9. The natural numbers form a set.
10. Every surjection has a right inverse.

These are Leinster's informal summary of Lawvere's ETCS axioms and Leinster spells them all out in categorical terms, just as we spelled several of them out already.

All these facts are familiar to every mathematician and are used routinely in textbooks. Munkres is fairly typical in suggesting that ZFC is the way to make them precise—though typical too in that he never actually names ZFC let alone states its axioms. And he repeatedly says the "precise" versions are not how mathematicians think. Lang is more artful in stating all and only the properties of sets that he actually uses, and never suggesting any others, are needed for rigor. All the properties Lang describes are easy consequences of these ETCS axioms.

The precise formal relationship between ETCS and ZFC has been known since Osius (1974). But, for a longer more practical/intuitive proof that ETCS suffices for standard mathematics, just read ordinary textbooks like those described here. Their proofs fit easily and naturally into the ETCS formalism.

Anyone who believes the ZFC axioms necessarily believes those of ETCS, but not conversely. The axioms of ETCS are all theorems of ZFC, when "function" is defined in the standard way for ZFC. The converse is not true.

The ZFC axioms say much more about sets than ETCS. These further claims are rarely noted, let alone used, in mathematics textbooks outside of set theory. They are pervasive in some kinds of advanced set theory. The ZFC axioms say everything is a set; notably functions are a special kind of set; and every set is built up from the empty set by transfinitely iterated set formation. We have seen how Munkres hints at these ZFC devices without going much into them and Lang avoids them. Section 1.5 will describe some reasons set theorists took them up, but first Section 1.4 describes why most textbooks don't.

1.4 Gauge Invariant Set Theory

Weatherall (2015) argues that the word "gauge" is ubiquitous in modern physics, and ambiguous. Without offering a thorough survey he describes two meanings, and we can use one of those:

> On the first strand, a "gauge theory" is a theory that exhibits excess structure . . . in such a way that (perhaps) one could remove some structure from the theory without affecting its descriptive or representational power. (Weatherall, 2015, 1–12)

As an example he gives electromagnetic theory using an electromagnetic potential A_a. Distinct potentials A_a, A'_a produce exactly the same observable consequences if their difference is the derivative of some scalar χ:

$$A_a - A'_a = \nabla_a \chi$$

In other words, given even ideal observations of the total behavior of some electromagnetic system, we have a choice of infinitely many different potentials A_a to describe that system.

So the specific potential is somehow more than we need. And in fact the potential is eliminated from versions of electromagnetic theory using electromagnetic force F_{ab} instead. Any two different forces $F_{ab} \neq F'_{ab}$ do produce, in principle, observably different consequences.

While mathematics has no good analogue of observable consequences, it certainly deals with descriptive or representational power. We will argue that the global membership relation of ZFC is a gauge on sets as it does not contribute to the descriptions and representations of structures in most of mathematics. This is not meant as any very close analogue to the situation in electromagnetic theory! But then, neither is the situation electromagnetic theory exactly the same as in General Relativity (Weatherall, 2015, passim).

We must distinguish between *global* and *local* membership. Throughout mathematics it is crucial to know which elements $x \in A$ of a set A are members of which subsets $S \subseteq A$. We say this relation is *local* to elements and subsets of the ambient set A. For example, arithmeticians need to know which natural numbers $n \in \mathbb{N}$ are in the subset of primes $Pr \subset \mathbb{N}$. On the other hand, nearly no one ever asks whether the imaginary unit $i \in \mathbb{C}$ is also a member of the unit sphere $S^2 \subset \mathbb{R}^3$, because they do not lie inside of any one natural ambient.

In ZFC the elements of sets are sets and it always makes sense to ask of sets X and Y whether $X \in Y$. This is a *global membership relation* since it does not rest on any sense of an ambient set but is meaningful for every two sets.

As a prominent example, it makes sense in ZFC to ask whether each rational number $q \in \mathbb{Q}$ is also a real number $q \in \mathbb{R}$, and the answer is "yes" or "no", depending on how we define real numbers. It is no if we use Dedekind cuts on the rational numbers, or equivalence classes of Cauchy sequences of rational numbers. It is yes if we use the definitions by Quine (1969, 136ff).

In ETCS membership is local, and it makes no sense to ask whether each rational number $q \in \mathbb{Q}$ is also a real number $q \in \mathbb{R}$ until both $\mathbb{Q}$ and $\mathbb{R}$ are taken as embedded in some common ambient set—and the most common ambient in textbook practice is $\mathbb{R}$ itself. In other words, ETCS takes it as true by stipulation that all rational numbers are real. It makes no sense in ETCS to ask for any answer except a stipulation. And most math texts that address the issue at all, including the two we looked at in Section 1.3, do solve it by stipulation. They say that every natural number is also a rational number, and every rational number is also a real number, without ever saying how this is to be achieved by a ZFC set-theoretic definition of those numbers. See McLarty (1993, 2008a) for more examples of textbook treatments of related issues and comparison with ETCS.

The global membership relation in ZFC is a gauge in the sense that it is rarely used in the "descriptive or representational" practices of mathematics, to recall the words of Weatherall (2015, 1–12) quoted earlier. Like the gauges in physical gauge theories it has important uses but the uses are of a technical kind which we will get to in the next section. Normal practice in most parts of mathematics describes and represents sets just up to isomorphism.[4]

Philosophers sometimes suppose definition up to isomorphism is a recent, abstract idea, which would not be found in classical mathematics. But this is backwards. Newton and Leibniz did understand the real numbers in terms of their algebraic and analytical relations to one another—which today are called isomorphism invariant—they did not understand real numbers as sets built up by transfinite accumulation from the empty set! They understood the arithmetic of whole numbers as dealing with the laws of addition and multiplication—which we today say define the integers only up to isomorphism—they never imagined one could explain what 2 *is* by using the set $\{\{\emptyset\}\}$ or the set $\{\emptyset, \{\emptyset\}\}$.

The novelty that arose in the nineteenth century, grew in the twentieth, and is growing faster today, was the scope, agility, and fecundity of structural methods.

For example, the idea of a tangent space to a manifold was reasonably clear, and extremely useful in expert hands, in the nineteenth century. But it became more useful and more widely accessible as the twentieth century gave clearer articulation to several distinct geometric constructions of it expressing different geometric aspects. One approach constructs it using coordinate systems placed on patches of the manifold. Another uses (equivalence classes of) curves through points of the manifold. Another uses (equivalence classes of) real-valued functions defined on neighborhoods of points of the manifold. Each of the different realizations relates most directly to some problems. All these constructions yield "the same structure" in some sense, which mathematicians at first described vaguely by saying all the realizations are "naturally isomorphic". That idea became more or less clear, to the more modern-minded mathematicians, through the early twentieth century. It was first addressed explicitly, and first made into a rigorous means of proving theorems, by Eilenberg and Mac Lane (1945) in the first general paper on categories and functors.

[4] Readers who compare Weatherall (2015) will note he associates gauges with a paucity of isomorphisms (so that some observationally identical models are not isomorphic), while we associate them with abundance of isomorphisms (so that gauges are not preserved by isomorphisms). That is because he takes 'isomorphism' in the way common throughout most of mathematics, where an isomorphism preserves the structures at hand. But isomorphisms in ZFC set theory do not preserve the membership structure. In ZFC or in ETCS, and indeed in Cantor's work long before either of those theories existed, an isomorphism of sets $f : A \to B$ is a function with an inverse, $f^{-1} : B \to A$, or in other words a one-to-one onto function. Such a function need not preserve the membership structure of ZFC sets (and almost never does, in fact). Given an isomorphism $f : A \to B$ and $z \in x \in A$ we can hardly conclude that $f(z) \in f(x)$. We cannot even conclude that $z \in A$, so we cannot conclude that $f(z)$ is defined at all. And if, in some particular case, $f(z)$ is defined, there is no reason for it to be an element of $f(x)$. This situation in ZFC is like the situation Weatherall (2015, 12) considers, where gauge transformations are taken as isomorphisms.

Philosophers often underestimate the practical need for rigorous proof in mathematics. There may be some quick, elementary proof of Fermat's Last Theorem that would avoid long difficult considerations—but honestly most people looking for proofs of Fermat's Last Theorem through the centuries since Fermat have looked for just that kind of proof. Most people working on Fermat's Last Theorem today continue to seek such a proof. Most have never learned the apparatus used by Wiles (1995). But up to now everyone who has found new and simpler proofs has used essentially that apparatus. All the yet-known proofs remain long and rather difficult, notably Kisin (2009).

Proofs like this require that each step be as clear and concise as possible while keeping extreme rigor. Otherwise, one could hardly credit the conclusion of such long reasoning. For many other examples, see McLarty (2008a, c).

Today mathematicians use extensive, efficient functorial means to deal with structures defined only up to isomorphism. And naturally, as we have seen in Lang and Munkres, they tend to treat sets also by isomorphism-invariant means. Why handle sets by one set of conceptual tools, and groups, rings, differential manifolds, and other algebaic or geometric structures by another, when the same tools will work for both? Those are the tools used in ETCS, and not in ZFC. But of course each part of mathematics also has its own special character which can call for specially adapted devices.

1.5 The Use of a Gauge

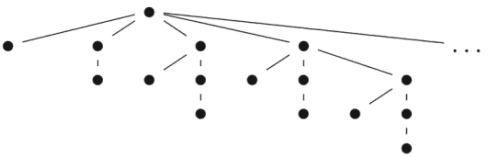
By and large mathematicians will agree with Munkres (2000, 36) that it is uninteresting to analyze a natural number $n \in \mathbb{N}$ or a real number $r \in \mathbb{R}$ as itself a set. But advanced investigations in set theory require much more detailed analysis of sets than most of mathematics does. Zermelo's approach to this was to give each set itself much more structure than Cantor had.[5] This attitude was canonized when the Axiom of Foundation and the Axiom Scheme of Replacement were added to Zermelo's early axiomatization to give today's Zermelo–Fraenkel Axioms. These axioms imply that every set is built up from the empty set by (possibly transfinitely) iterated set formation.

In other words every ZFC set has a well-founded downward-growing *membership tree* structure where the top node indicates the set, and each node has as many nodes below it as elements. For example, the set $\mathbb{N} = \{0, 1, 2, 3, \dots\}$ can be represented by the set of finite von Neumann ordinals:

$$\{\emptyset,\ \{\emptyset\},\ \{\emptyset, \{\emptyset\}\},\ \{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\},\ \dots\}.$$

[5] Zermelo declared his preference for Frege on this question, over Cantor, in his comments on Cantor's collected works. See especially Cantor (1932, 351, 353, 441f.) and Lawvere (1994).

So it has this tree:



Each bottom node stands for the empty set ∅, each node with nothing below it but a single bottom node stands for the singleton set {∅}, and so on. Each branch is finite but they get longer and longer without bound as they go to the right.

Now be clear, both ETCS and ZFC can describe tree structures like this. The difference is that ZFC says each set *S* has such a tree intrinsically given by the membership relation on its elements, and their elements, and so on. Because the elements of *S* are ZFC sets, so are their elements, and so on. So ZFC sets come intrinsically arranged in an *cumulative hierarchy* also called the *iterative hierarchy* where each set is located at the lowest stage higher than any of its elements.

This tree structure is extremely useful in set theory. The *Mostowski embedding theorem*, found in any advanced ZFC textbook, tells exactly which trees correspond to ZFC sets. And categorical set theory also uses it for logical investigations (Osius, 1974, 88). A *tree interpretation* lets ETCS interpret ZFC with no loss of information at all, by interpreting a ZFC set as a suitable tree as discussed in McLarty (2004). So there is a translation routine taking any theorem or proof in ZFC to an equivalent theorem or proof in ETCS.[6]

In short, what ZFC takes as a set *S*, ETCS takes as a set *S* plus a suitable (extensional, well-founded) tree with the elements of *S* as the first level of nodes below the top. The tree structure in ETCS captures the iterative hierarchy structure in ZFC. From the viewpoint of ETCS the iterative hierarchy is a gauge, very useful in set theory research.

Then there is a yet further structure, called *constructibility*, forming the *constructible hierarchy*, with similarly pervasive research uses. A constructible set is not only accumulated level by level from earlier constructible sets, but each set is formed with an explicitly stated definition of which earlier sets are to be collected into it.[7] This gives an extremely tight handle on each set and it is useful at every level of research set theory. The set theories closest to arithmetic generally assume every set is constructible to gain enough control on the sets, as explained by Simpson (2010, 282).

[6] The ETCS version may need to specify further assumptions corresponding to the fact that ZFC assumes stronger axioms than ETCS. The corresponding stronger assumptions are always available in ETCS, if only by translating the ZFC version, though in most natural cases there is also a natural correspondent in ETCS.

[7] Actually, on conventional accounts, each constructible set *S* is formed infinitely many times, each time with a different explicitly stated definition. But the definitions are well ordered so each constructible *S* has a unique *first* definition forming it.

However, nearly the whole theory of large cardinals disappears if all sets are constructible. *Measurable cardinals* cannot exist if all sets are constructible, and so of course no larger cardinals can either. So ZFC set theorists treat constructibility as a valuable feature of *some* sets, not *all* sets. But note this is not a distinction at the level of set isomorphism. The ZFC axioms themselves trivially prove every set is *isomorphic to* some constructible set: choice proves every set is isomorphic to some ordinal. All ordinals are constructible by definition.

In short, ZFC and ETCS both treat membership trees and constructibility as gauges. Both traits are extremely useful at times (as are potentials in electromagnetism). Neither trait is preserved by set isomorphisms. Neither trait is much mentioned in math outside of research set theory. These gauges are indispensable to the kind of questions handled in set theory texts such as Kunen (1983) or Kanamori (1994). These texts find it natural to use ZFC, which has the first gauge (membership trees) built in, and which does not imply that every set admits the second gauge at all. But far the greatest part of mathematics in algebra, geometry, or analysis is gauge invariant—or, as logicians prefer to say, isomorphism invariant—both theoretically and in daily practice.

References

Artin, M., Grothendieck, A., and Verdier, J.-L. (1972). *Théorie des Topos et Cohomologie Etale des Schémas*. Séminaire de géométrie algébrique du Bois-Marie, 4. Springer-Verlag, Paris. Three volumes, usually cited as SGA 4.

Bourbaki, N. (1949). Foundations of mathematics for the working mathematician. *Journal of Symbolic Logic* 14, 1–8.

Bourbaki, N. (1958). *Théorie des Ensembles*, 3rd edition. Hermann, Paris.

Cantor, G. (1932). *Gesammelte Abhandlungen mathematischen und philosophischen Inhalts*, ed. E. Zermelo. Springer, Berlin.

Cohen, P. (1966). *Set Theory and the Continuum Hypothesis*. W. A. Benjamin, New York.

Eilenberg, S., and Mac Lane, S. (1945). General theory of natural equivalences. *Transactions of the American Mathematical Society* 58, 231–94.

Fantechi, B., Vistoli, A., Gottsche, L., Kleiman, S. L., Illusie, L., and Nitsure, N. (2005). *Fundamental Algebraic Geometry: Grothendieck's FGA Explained*, vol. 123 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI.

Freitag, E., and Kiehl, R. (1988). *Étale cohomology and the Weil conjecture*. Springer-Verlag, New York.

Gödel, K. (1939). The consistency of the axiom of choice and of the generalized continuum-hypothesis. *Proceedings of the National Academy of Sciences of the United States of America* 24, 556–7.

Hájek, P., and Pudlák, P. (1993). *Metamathemtics of First Order Arithmetic*. Springer-Verlag, New York.

Kanamori, A. (1994). *The Higher Infinite*. Springer-Verlag, New York.

Kelley, J. (1955). *General Topology*. Van Nostrand, New York.

Kisin, M. (2009). Moduli of finite flat group schemes, and modularity. *Annals of Mathematics* 170(3), 1085–180.

Kreisel, G. (1987). So-called formal reasoning and the foundational ideal, in Paul Weingartner and Gerhard Schurz (eds), *Logik, Wissenschaftstheorie und Erkenntnistheorie*, 19–42. Hölder-Pichler-Tempsky, Wien.

Kunen, K. (1983). *Set Theory: An Introduction to Independence Proofs*. North-Holland, Amsterdam.

Lang, S. (1965). *Algebra*, 1st edn. Addison-Wesley, Reading, MA.

Lang, S. (1993). *Algebra*, 3rd edn. Addison-Wesley, Reading, MA.

Lang, S. (2005). *Algebra*. Addison-Wesley, Reading, MA.

Lawvere, F. W. (1964). An elementary theory of the category of sets. *Proceedings of the National Academy of Science of the United States of America* 52, 1506–11.

Lawvere, F. W. (1994). Cohesive toposes and Cantor's "lauter Einsen". *Philosophia Mathematica* 2, 5–15.

Leinster, T. (2014). Rethinking set theory. *American Mathematical Monthly* 121(5), 403–15.

Linnebo, Ø., and Pettigrew, R. (2011). Category theory as an autonomous foundation. *Philosophia Mathematica* 19, 227–54.

Lurie, J. (2009). *Higher Topos Theory*. Annals of Mathematics Studies 170. Princeton University Press, Princeton, NJ.

McLarty, C. (1990). The uses and abuses of the history of topos theory. *British Journal for the Philosophy of Science* 41, 351–75.

McLarty, C. (1993). Numbers can be just what they have to. *Noûs* 27, 487–98.

McLarty, C. (2004). Exploring categorical structuralism. *Philosophia Mathematica*, 37–53.

McLarty, C. (2008a). What structuralism achieves, in Paolo Mancosu (ed.), *The Philosophy of Mathematical Practice*, 354–69. Oxford University Press, Oxford.

McLarty, C. (2008b). Articles on Claude Chevalley, Jean Dieudonné, and André Weil, in Noretta Koertge (eds), *New Dictionary of Scientific Biography*, vol. II, 116–20 and 289–93, 254–8. Scribner's, Detroit.

McLarty, C. (2008c). "There is no ontology here": visual and structural geometry in arithmetic, in Paolo Mancosu (ed.), *The Philosophy of Mathematical Practice*, 370–406. Oxford University Press, Oxford.

McLarty, C. (2010). What does it take to prove Fermat's Last Theorem? *Bulletin of Symbolic Logic* 16, 359–77.

McLarty, C. (2011). A finite order arithmetic foundation for cohomology. Preprint on the mathematics arXiv, arXiv:1102.1773v3, 2011.

McLarty, C. (2012). Categorical foundations and mathematical practice. *Philosophia Mathematica* 20, 111–13.

Milne, J. (1980). *Étale Cohomology*. Princeton University Press, Princeton, NJ.

Munkres, J. (2000). *Topology*, 2nd edn. Prentice Hall, Upper Saddle River, NJ.

Osius, G. (1974). Categorical set theory: A characterization of the category of sets. *Journal of Pure and Applied Algebra* 4, 79–119.

Quine, W. V. O. (1969). *Set Theory and Its Logic*. Harvard University Press, Cambridge, MA.

Rudin, M. E. (1972). The box product of countably many compact metric spaces. *General Topology and its Applications* 2, 293–8.

Russell, B. (1924). Logical atomism, in J. Muirhead (ed.), *Contemporary British Philosophers*, 356–83. Allen and Unwin, London. Reprinted in Russell, B. (1956) *Logic and Knowledge*, 323–43. Allen and Unwin, London.

Simpson, S. (2010). *Subsystems of Second Order Arithmetic*. Cambridge University Press, Cambridge, UK.

Spivak, M. (1999). *A Comprehensive Introduction to Differential Geometry*, 3rd edn. Publish or Perish, Houston.

Tamme, G. (1994). *Introduction to Etale Cohomology*. Springer-Verlag, New York.

Weatherall, J. (2015). Understanding gauge. Preprint on the mathematics arXiv, arXiv: 1505.02229.

Wiles, A. (1995). Modular elliptic curves and Fermat's Last Theorem. *Annals of Mathematics* 141, 443–551.

2

Reviving the Philosophy of Geometry

David Corfield

2.1 Introduction

Leafing through Robert Torretti's book *Philosophy of Geometry from Riemann to Poincaré* (1978), it is natural to wonder why, at least in the Anglophone community, we have little activity meriting this name today. Broadly speaking, we can say that any philosophical interest in geometry shown here is directed at the appearance of geometric constructions in physics, without any thought being given to the conceptual development of the subject within mathematics itself. This is in part a result of a conception we owe to the Vienna Circle and their Berlin colleagues that one should sharply distinguish between *mathematical* geometry and *physical* geometry. Inspired by Einstein's relativity theory, this account, due to Schlick and Reichenbach, takes mathematical geometry to be the study of the logical consequences of certain Hilbertian axiomatizations. For its application in physics, in addition to a mathematical geometric theory, one needs laws of physics and then 'coordinating principles' which relate these laws to empirical observations. From this viewpoint, the mathematics itself fades from view as a more or less convenient choice of language in which to express a physical theory. No interest is taken in which axiomatic theories deserve the epithet 'geometric'.

However, in the 1920s this view of geometry did not go unchallenged as Hermann Weyl, similarly inspired by relativity theory, was led to very different conclusions. His attempted unification of electromagnetism with relativity theory of 1918 was the product of a coherent geometric, physical, and philosophical vision, inspired by his knowledge of the works of Fichte and Husserl. While this unification was not directly successful, it gave rise to modern gauge field theory. Weyl, of course, also went on to make a considerable contribution to quantum theory. And while Einstein gave initial support to Moritz Schlick's account of his theory, he later became an advocate of the idea that mathematics provides important conceptual frameworks in which to do physics:

> Experience can of course guide us in our choice of serviceable mathematical concepts; it cannot possibly be the source from which they are derived; experience of course remains the sole criterion of the serviceability of a mathematical construction for physics, but the truly creative principle resides in mathematics. (Einstein, 1934, 167)

We may imagine then that an important chapter in any sequel to Torretti's book would describe both Reichenbach's and Weyl's views on geometry. This is done in Thomas Ryckman's excellent *The Reign of Relativity* (2005), where the author also discusses further overlooked German-language philosophical writings on geometry from the 1920s, this time by Ernst Cassirer. Cassirer's extraordinary ability to assimilate the findings of a wide range of disciplines sees him discuss the work of important mathematicians such as Felix Klein, Steiner, Dedekind, and Hilbert.

Ryckman ends his book with a call for philosophical inquiry into what sense a 'geometrized physics' can have, to emulate the work of these thinkers from the interwar period. And he is not alone in thinking that this was a golden age. There is now an impressive concentration on this era. Today, in Ryckman and similar-minded thinkers, such as Michael Friedman and Alan Richardson (see the contributions of all three in Domski and Dixon (2010)), we find fascinating discussions about these themes. However, these discussions by themselves are unlikely to give rise today to the kind of primary work that they are studying from the past. It is one thing to make a careful, detailed study of the interweaving of philosophy, mathematics, and physics of a period, quite another to begin to take the steps necessary for a revival of such activity.

Michael Friedman (2001) points out two connected, yet somewhat distinct, activities dealing with mathematical physics which might be called 'philosophical'. One termed 'meta-scientific' is much as Weyl does, reconceiving the idea of space and thereby generating foundational advances. Meta-science is typically done by philosophically informed scientists, such as Riemann, Helmholtz, Poincaré, and Einstein. By contrast, the other activity is much as Cassirer and the Vienna Circle did, reflecting on the broader questions of the place of mathematics and science in our body of knowledge in light of important events in the histories of those practices. While there spontaneously arises work of the first kind in any era, work of the second kind requires a philosophical orientation which may be lost. One very obvious difference is that today we have so few philosophers emulating Cassirer by keeping abreast of the mathematics of the recent past. This simply must change if we are to generate the forms of discussion to parallel those of the 1920s. For too long, philosophy has thought to constrain its interest in any current mathematical research largely to set theory, when it has long been evident that it offers little or nothing as far as many core areas of mathematics are concerned, and especially the mathematics needed for physics. Casting the differential cohomology of modern quantum gauge theory in set-theoretic clothing would do no favours to anyone. So, with some notable exceptions, such as

Marquis (2008) and McLarty (2008), we let the bulk of mainstream mathematical research pass us by.

However, there are reasons to be hopeful. I shall argue in this chapter that our best hope in reviving a 1920s-style philosophy of geometry lies in following what has been happening at the cutting edge of mathematical geometry over the past few decades, and that while this may appear a daunting prospect, we do now have ready to hand a means to catch up rapidly. These means are provided by what is known as *cohesive homotopy type theory*.

Univalent foundations, or plain homotopy type theory (cf. Awodey's and Shulman's chapters), provide the syntax for theories which can be interpreted within $(\infty,1)$-toposes, a generalization of the ordinary notion of toposes. The basic shapes of mathematics are now taken to be the so-called 'homotopy n-types'. However, these are not sufficient to do what needs to be done in modern geometry, and especially in the geometry necessary for modern physics, since we need to add further structures to express continuity, smoothness, and so on. As we add extra properties and structures to $(\infty,1)$-toposes, characterized by qualifiers—local, ∞-connected, cohesive, differentially cohesive—increasing amounts of mathematical structure are made possible internally. The work of Urs Schreiber (2013) has shown that cohesive $(\infty,1)$-toposes provide an excellent environment to approach Hilbert's sixth problem on axiomatizing physics, allowing the formulation of relativity theory and all quantum gauge theories, including the higher-dimensional ones occurring in string theory.

Cohesiveness in this sense arose from earlier formulations of the notion in the case of ordinary toposes by William Lawvere (2007), motivated in turn by philosophical reflection on geometry and physics. Schreiber's claim, however, is that for these concepts to take on their full power they must be extended to the context of higher-topos theory, that is the theory of $(\infty,1)$-toposes, where differential cohomology finds its natural setting. Now, rather than the mathematics necessary for physics being viewed, as it often is at present from a set-theoretic foundation, as elaborate and unprincipled, we can see the simplicity of the necessary constructions through the universal constructions of higher-category theory.

In a single article it will only be possible to outline the kind of work necessary to fill in the spaces we have left ourselves. There is important interpretative work to do already in making sense of plain homotopy type theory, so here I can only indicate further work to be done. At the same time, in that mathematics finds itself once again undergoing enormous transformations in its basic self-understanding, it is important as philosophers to take this opportunity to remind ourselves that we should provide an account of mathematical enquiry where such changes are to be expected. It is striking that Hegel should be found informing both those wishing to characterize the dynamic growth of mathematics and those striving to refashion the very concepts of modern geometry itself.

2.2 Current Geometry

If any reassurance is needed that geometry is alive and well today, one need only look at the variety of branches of mathematics bearing that name which are actively being explored:

algebraic, differential, metric, symplectic, contact, parabolic, convex, Diophantine, tropical, conformal, Riemannian, Kähler, Arakelov, analytic, rigid analytic, global analytic, . . .

Continuing our search, we find noncommutative versions of some of these items, also 'derived' versions, and so on. Indeed, there has never been so much 'geometric' research being carried out as there is today, from constructions that Gauss or Riemann might have recognized to ones which would seem quite foreign. So the question arises of whether this list provides just a motley of topics which happen to bear the same name, or whether there is something substantial that is common to all of them, or at least many of them.

Evidence for the latter option comes from people still making unqualified use of 'geometry' and its cognates to mean something. Some such uses are informal, as in the following example:

> The fundamental aims of geometric representation theory are to uncover the deeper geometric and categorical structures underlying the familiar objects of representation theory and harmonic analysis, and to apply the resulting insights to the resolution of classical problems.
>
> (MRSI, 2014)

Other uses are technical, such as where Jacob Lurie (2009) uses the term 'geometry' to name a certain kind of mathematical entity, here a small $(\infty, 1)$-category with certain additional data. The question then arises as to what features of these structures make Lurie single them out as geometries. At first glance this seems a rather technical matter. Let us return to it once we have some motivation from the past.

Something that would have seemed novel to Gauss and Riemann, and which might give rise to doubts concerning the unity of geometry, is the thorough injection of spatial ideas brought into algebraic geometry by Alexandre Grothendieck in the 1960s. By the late 1800s it was already known of the collection of complex polynomials, $\mathbb{C}[z]$, that the space on which these functions are defined could be recaptured from the algebraic structure of the collection itself. $\mathbb{C}[z]$ forms a ring, and it is possible to construct an associated space from points corresponding to its maximal ideals. These are the ideals generated by $(z-a)$, for each $a \in \mathbb{C}$. Picking up on the complex function field/algebraic number field analogy of Dedekind and Weber, as developed by Weil in his Rosetta Stone account (see Corfield, 2003, ch. 4), it was then shown that even apparently nonfunctional rings, such as the ring of integers and others encountered in arithmetic, might be treated likewise. Grothendieck's scheme theory (see McLarty, 2008) provided such a space, in the case of the integers denoted $Spec(\mathbb{Z})$, again constructed out of prime ideals. Now an integer is considered as a function defined at each prime, a point

in $Spec(\mathbb{Z})$, as the function $n(p) \equiv n \,(mod\, p)$. Where this differs from the complex function case is that while the values of 'integer-as-function' still land in a field, here the field varies according to the point where the function is evaluated, $\mathbb{F}_p$ as p varies. This suggests a space whose points are not identical.

This attempt to geometrize arithmetic is not an empty game. It feeds through to mathematical practice as can readily be discovered thanks to the growth of online informal discussion.

I like to picture $Spec\,\mathbb{Q}$ as something like a 2-manifold which has had all its points deleted. The extra complication is that what we think of as the points are actually very small circles. So it's really a three manifold with all of the loops inside it deleted.

For example, let's look first at function fields. $Spec\,\mathbb{C}[z]$ is just the complex line $\mathbb{C}$. As we start inverting elements of $Spec\,\mathbb{C}[z]$, as we must do to make $Spec\,\mathbb{C}(z)$, the effect on the spectrum is to remove bigger and bigger finite sets of points. The limit is where we remove all the points and we're just left with some kind of mesh.

If we had started with a Riemann surface of genus g, then we'd be left with a mesh of genus g, a surface sewn out of the cloth from which fly screens for windows are made. If we want to recover the original surface from the surface mesh, we just put it out back in the shed for a while and let the mesh fill up with dirt. This is just the familiar fact that a (smooth compact, say) Riemann surface can be recovered from the field of meromorphic functions on it.

If we replace $\mathbb{C}$ by a finite field $\mathbb{F}$, then everything is the same but what we thought of as the point is now a very small circle, and so our original surface reveals itself to be a 3-manifold fibered over a very small circle when we zoom in. And when we delete points, we're really deleting not just single-valued sections of this fibration but also multivalued sections. So $Spec\,\mathbb{F}[z]$ is some kind of 3-manifold fibered over the circle with all the loops over the base circle deleted.

For the passage from $\mathbb{Z}$ to $\mathbb{Q}$, I don't have anything better to say than that it's sort of the same but there's no base circle. We're just removing lots of loops from a 3-manifold. Maybe some should be seen as bigger than others, corresponding to the fact that there are prime numbers of different magnitudes. (Borger, 2009)

Here we see Borger passing across each of the three columns of Weil's Rosetta Stone.

Now, not only do we find a geometrized arithmetic, but these ideas and constructions are the structural cousins of those appearing in cutting-edge physics, as we see in this comment by David Ben-Zvi:

the geometric analog of a number field or function field in finite characteristic should not be a Riemann surface, but roughly a surface bundle over the circle. This explains the "categorification" (need for a function-sheaf dictionary, which is the weak part of the analogy) that takes place in passing from classical to geometric Langlands—if you study the corresponding QFT on such three-manifolds, you get structures much closer to those of the classical Langlands correspondence. (Ben-Zvi, 2014)

So we have *both* widespread current interest in classically geometric areas of mathematics and geometric approaches to other areas, including arithmetic, *and* ways of thinking about the subject matter expressed, at least informally, in a very visual

language. Geometry as a whole is something larger that that which has application in mathematical physics, and applied mathematics more generally. Similar structures are now found to lie at the heart of number theory. But how to go about saying something satisfactorily general about geometry?

One might throw up one's hands at the task of bringing this wealth of subject matter under the umbrella of a straightforward description. To the extent that people try to do this it is largely left to the doyens of mathematics. For example, Sir Michael Atiyah writes

> Broadly speaking, I want to suggest that geometry is that part of mathematics in which visual thought is dominant whereas algebra is that part in which sequential thought is dominant.
> (Atiyah, 2003, 29)

Such a distinction is reminiscent of Kant, for whom space and time were considered to be forms of sensibility, and yet Atiyah continues:

> This dichotomy is perhaps better conveyed by the words "insight" versus "rigour" and both play an essential role in real mathematical problems. (Atiyah, 2003, 29)

A more careful treatment is required here, since there seems nothing to object to in the idea of 'rigorous geometry' or 'algebraic insight'. We need to turn back the clock to when philosophical research directed itself towards then current geometry.

2.3 Regaining the Philosophy of Geometry

What led to the demise of the philosophy of geometry in the English-speaking world? I think this can be attributed largely to the success of logical empiricism. Many of those dispersed from Germany and Austria in the 1930s were accepted into the universities of the USA, welcomed by existing empiricists such as Ernst Nagel. In a long paper published in 1939, Nagel uses the history of projective geometry to explain the new understanding of then modern axiomatic mathematics:

> It is a fair if somewhat crude summary of the history of geometry since 1800 to say that it has led from the view that geometry is the apodeictic science of space to the conception that geometry, in so far as it is part of natural science, is a system of "conventions" or "definitions" for ordering and measuring bodies. (Nagel, 1939, 143)
>
> The distinction between a pure and an applied mathematics and logic has become essential for any adequate understanding of the procedures and conclusions of the natural sciences.
> (Nagel, 1939, 217)

So axiom systems are proposals for stipulations. As pure mathematics they are to be studied for their logical properties. Some of them may be found to be well adapted to allow the expression of scientific laws, which may then be used in applied sciences. This is made possible by coordinating principles which tie the scientific laws to empirical measurements. For example, Riemannian geometry allows for the

expression of Einstein's field equations, which can be coordinated to observation by stipulating that light follows null geodesics.

While now the ideational content of mathematics is left to one side, other tasks fall to the philosopher of mathematics:

> the concepts of structure, isomorphism, and invariance, which have been fashioned out of the materials to which the principle of duality is relevant, dominate research in mathematics, logic, and the sciences of nature. (Nagel, 1939, 217)

Had philosophers at least heeded this, more attention might have been paid to category theory, the language par excellence of structure, isomorphism, and invariance, which emerged shortly after Nagel's paper (Corfield, 2015). As it was, a uniform treatment of mathematics as the logical consequences of definitions, or of the set-theoretic axioms, came to prevail.

In the process, as Heis (2011) argues, two lines of thought from earlier in the century were being ruled out, each responding to other nineteenth-century developments in geometry:

1. What could be saved of Kantian philosophy given the appearance of non-Euclidean geometry, and then Riemannian geometry? What are the conditions for spatial experience?
2. How should we understand the ever-changing field of geometry given the introduction of ideal elements, imaginary points, and so on?

Let us take each of these in turn.

2.3.1 Weyl: the essence of space

With an ever-expanding variety of geometries emerging through the nineteenth century, it became implausible to maintain with Kant that our knowledge of Euclidean geometry is a priori. Helmholtz had argued that because empirical measurement requires that objects undergo only 'rigid motions', we can work out which geometries are presupposed by our physics. He concluded that only those spaces which possessed the property of constant curvature were permissible. With the contribution of the technical expertise of Sophus Lie, this line of research resulted in the Helmholtz–Lie theorem, characterizing Euclidean, elliptic, and hyperbolic geometries.

Research such as this was certainly discussed by philosophers. Indeed, Russell, and later Schlick and Reichenbach, responded to Helmholtz's work, but perhaps the most profound response came from Hermann Weyl. After the success of Einstein's general theory of relativity, Helmholtz's results were evidently far too limited. Weyl, inspired by Husserl and perhaps more profoundly by Fichte (see Scholz, 2005), sought to discern what he termed "the essence of space". In a letter to Husserl, he wrote

> Recently, I have occupied myself with grasping the essence of space [das Wesen des Raumes] upon the ultimate grounds susceptible to mathematical analysis. The problem accordingly

concerns a similar group theoretical investigation, as carried out by Helmholtz in his time (Ryckman, 2005, 113)

Weyl conceived of spaces in which it was only possible to compare the lengths of two rods if they were situated at the same point.

> Only the spatio-temporally coinciding and the immediate spatial-temporal neighborhood have a directly clear meaning exhibited in intuition.
>
> (Weyl, 'Geometrie und Physik', quoted in Ryckman (2005, 148))

Along the lines of Helmholtz and Lie, this led him to prove a group theoretic result: the only groups satisfying certain desiderata (involving the "widest conceivable range of possible congruence transfers" and a demand for a single affine connection) are the special orthogonal groups of any signature with similarities, $G \simeq SO(p, q) \times \mathbb{R}^+$ (Scholz, 2011, 230).

There is an interesting story to be told here of how Einstein and other physicists found implausible the possibility allowed by this geometry that rods of identical lengths, as measured at one point, if transported along different paths to a distant point might have different resulting lengths. One usually tells the story of how the beauty of the mathematics got the better of Weyl, and how physicists eventually uncovered what was good about the idea whilst modifying his original idea to allow a $U(1)$ gauge group. This story needs to be told in a much more nuanced way (see Giovanelli, 2013), and in any case is complicated by the survival of Weyl's original idea in forms of conformal gauge theory.

In any case, Weyl himself later became sceptical of this kind of mathematical speculation about the geometry required for physics that had so consumed him in his earlier years. With the demise of other philosophical attempts to study our a priori geometric intuition, for example, Carnap's doctoral thesis on how our intuitive concept of space was required to be n-dimensional topological space, such attempts largely came to an end. We will take a look at more recent 'meta-scientific' kinds of work, but first let us turn to Friedman's other form of philosophical research.

2.3.2 Cassirer: beyond intuition

In an unusual paper, published the year before his death, Ernst Cassirer (1944) argued for an important connection to be seen between Felix Klein's Erlanger programme and our everyday perception. Where Klein had given a presentation of many forms of geometry as the study of invariants of space under the action of groups of symmetry, Cassirer saw the seeds for this idea in our abilities to perceive the invariant size, colour, and shape of objects under varying viewing conditions. Now evidently these abilities are rooted in our distant evolutionary past, and yet the full-blown mathematical idea had only crystallized in modern mathematical thinking around 1870.

Klein's ideas on geometry marked an important stage in the course of a revolutionary century for geometric thought. Not only had the range of geometries been extended from the single Euclidean geometry to hyperbolic and elliptical forms, but

there had been many kinds of extension of the notion of space by the introduction of elements that seemed to lead us away from the intuitively familiar. For example, the nondegenerate conic sections had been unified as curves of degree 2 in complex projective space, brought about by the addition of 'points at infinity' and complex coordinates. Now all circles were seen to pass through two imaginary points at infinity, and so 'intersect' there.

Such forays beyond the intuitive led those wishing to retain what they took to be valuable in Kant to take a different tack. As Heis (2011) convincingly shows, the neo-Kantian Cassirer had to come to terms with just such developments. This is evident in his later work:

> It is hence obvious that mathematical theories have developed in spite of the limits within which a certain psychological theory of the concept tried to confine them. Mathematical theory ascended higher and higher in order to look farther and farther. Again and again it ventured the Icarian flight which carried it into the realm of mere "abstraction" beyond whatever may be given and represented in intuition. (Cassirer, 1944, 24)

But then without any firm rootedness in intuition, what provides us with guidance that our "Icarian flights" are heading in the best direction? This problematic runs through Cassirer's career, and is answered by the unity of the history of the discipline.

> Though a properly Neo-Kantian philosophy of mathematics will appreciate that mathematics itself has undergone fundamental conceptual changes throughout its history, such a philosophy will also have to substantiate the claim that the various stages in the historical development of mathematics constitute *one history* . . . we can say that they [mathematicians] were studying the same objects only because we can say that they are parts of the same history. (Heis, 2011, 768)

It is worth quoting Cassirer at length on this point:

> it is not enough that the new elements should prove equally justified with the old, in the sense that the two can enter into a connection that is free from contradiction—it is not enough that the new should take their place beside the old and assert themselves in juxtaposition. This merely formal combinability would not in itself provide a guarantee for a true inner conjunction, for a *homogeneous logical structure of mathematics*. Such a structure is secured only if we show that the new elements are not simply adjoined to the old ones as elements of a different kind and origin, but the new are *a systematically necessary unfolding of the old*. And this requires that we demonstrate a primary logical kinship between the two. Then the new elements will bring nothing to the old, other than *what was implicit in their original meaning*. If this is so, we may expect that the new elements, instead of fundamentally changing this meaning and *replacing* it, will first bring it to *its full development* and *clarification*. (Cassirer, 1957, 392)

If one can hear an overtone of Hegelian thought here, this is not surprising. In the introduction to this third volume of *The Philosophy of Symbolic Forms* Cassirer explained the debt to Hegel as shown by the subtitle of the book—*The Phenomenology of Knowledge*:

> The truth is the whole—yet this whole cannot be presented all at once but must be unfolded progressively by thought in its own autonomous movement and rhythm. It is this unfolding which constitutes the being and essence of science. The element of thought, in which science is and lives, is consequently fulfilled and made intelligible only through the movement of its becoming. (Cassirer, 1957, xiv)

This line of thought sits very happily with the idea that important developments in a discipline allow its history to be written in such a way that it makes best sense of what was only obscurely seen in the past or of what became the means to overcome perceived obstacles or limitations, in other words, a history of rational unfolding out of an older stage. As I argue in Corfield (2012), we find this position very well expressed by the moral philosopher Alasdair MacIntyre. A similar idea is expressed by Friedman's 'retrospective' rationality (2001).

In the Anglophone revival (Friedman, Ryckman, Richardson, Heis, and Everett) of interest in Cassirer of recent years, there has been particular focus on the place of the 'constitutive' and the 'regulative' in his account of the progress of science. This amounts to rival interpretations of the relative importance for Cassirer of the prospective overcoming of limitations within a discipline and the retrospective rationalization of its course, eventually as seen from an ideal future point. However these debates turn out, it is intriguing then to see what we might call a further strand added in the 1944 paper, that in mathematics we may devise concepts which owe their origin to unnoticed cognitive structures. In the case of the Erlanger Program

> the mathematical concepts are only the full actualisation of an achievement that, in a rudimentary form, appears also in perception. Perception too involves a certain invariance and depends upon it for its inner constitution. (Cassirer, 1944, 17)

Taken together, we can see in the work of Weyl and Cassirer just how far we are here in attitude towards mathematical geometry from what was bequeathed to us by the logical empiricists. Heis quotes Hans Reichenbach:

> It has become customary to reduce a controversy about the logical status of mathematics to a controversy about the logical status of the axioms. Nowadays one can hardly speak of a controversy any longer. The problem of the axioms of mathematics was solved by the discovery that they are definitions, that is, arbitrary stipulations which are neither true nor false, and that only the logical properties of a system—its consistency, independence, uniqueness, and completeness—can be subjects of critical investigation. (Heis, 2011, 790)

He very aptly writes: "One could hardly find a point of view further from Cassirer's own" (Heis, 2011, 790). Indeed so, but now to be true to Cassirer's spirit we should try to work out our own position on the rationality of mathematical enquiry in the process of coming to frame what has been happening in the mathematics of the recent past. That this has typically not been felt to be a requirement of philosophy makes this no easy task, but we should try to make a start anyway.

2.4 Capturing Modern Geometry

Even to summarize one particular line of development here will not be easy. Along with Grothendieck's invention of scheme theory, mentioned in Section 2.2, we would also need to talk about topos theory. This could take the form of a story of natural unfolding. Indeed, we could report the originator's own words.

> one can say that the notion of a topos arose naturally from the perspective of sheaves in topology, and constitutes a substantial broadening of the notion of a topological space, encompassing many concepts that were once not seen as part of topological intuition . . . As the term "topos" itself is specifically intended to suggest, it seems reasonable and legitimate to the authors of this seminar to consider the aim of topology to be the study of topoi (not only topological spaces).
>
> (Grothendieck and Verdier, 1972, 302)

However, for the purposes of this chapter, we need to race forward to much more recent work. Jacob Lurie motivates his 'Structured Spaces' paper (2009) by means of an account of the passage to less restricted forms of Bézout's theorem. This is a result that goes back to the eighteenth century, involving just the kind of achievement of unity through addition of ideal elements that interested Cassirer. While it was known to Newton that the number of real solutions to the intersection of a pair of plane curves was bounded by the product of their degrees, by the nineteenth century we find a form of the result that states that two complex projective plane curves of respective degrees m and n which share no common component have $m \cdot n$ points of intersection, counted with multiplicity. Any two non identical conics meet four times, including at those imaginary points at infinity in the case of two circles that were mentioned in the previous section.

Lurie takes this result up, looking to understand it in terms of cohomology and the cup product of fundamental classes of the curves, which corresponds to the class of their intersection. Since this method does not work for non-transverse intersections, using Grothendieck's constructions we then turn to 'nonreduced' schemes. Further, according to Lurie, we should look at a Euler characteristic involving the dimension of the local ring of the scheme-theoretic intersection plus various corrections.

Now an interesting thing happens when we attempt to retain the fundamental result $[C] \cup [C'] = [C \cap C']$ in the very general setting where there may even be coinciding components. Here we need *derived* algebraic geometry.

> To obtain the theory we are looking for, we need a notion of generalized ring which remembers not only whether or not x is equal to 0, but how many different ways x is equal to 0. One way to obtain such a formalism is by categorifying the notion of a commutative ring. That is, in place of ordinary commutative rings, we consider categories equipped with addition and multiplication operations (which are encoded by functors, rather than ordinary functions). (Lurie, 2009, 3)

Lurie is drawing attention here to the passage from the proposition 'x is equal to 0' to the set of ways in which it is equal. To do so is to take the first step up an infinitely tall ladder of weakenings of identity. In Corfield (2003, ch. 10) I give an account of categorification, the replacement of set by category by 2-category. In the dozen or so

years since my book, the emphasis has swung round to the groupoid version of this ladder, where sets become groupoids become 2-groupoids, etc. Lurie, in particular, was instrumental in this change in showing that most constructions of ordinary category theory have their analogues in the $(\infty, 1)$ setting, where instead of hom-sets between objects, we have ∞-groupoids.

Now, since when dealing with a pair of coinciding lines, we need to make identifications in the form of isomorphisms, we find the following:

> These isomorphisms are (in general) distinct from one another, so that the categorical ring C "knows" how many times x and y have been identified. (Lurie, 2009, 4)

Of course, we never stop with a single step up this ladder, and eventually we seek further generalized forms of ring, such as E_∞-ring spectra and simplicial commutative rings. The key lesson here is that to retain a simple formulation, we must change our framework, for one thing here to allow homotopic weakening. In Cassirerian terms, this is forced upon us by the natural unfolding of the discipline. And it is not just algebraic geometry that demands this richer notion of space, so does physics. The moduli spaces of today's gauge field theories are often stacks, such as the moduli stack of flat connections for some gauge group. Higher gauge theory requires similar homotopic weakening to higher stacks (Schreiber, 2013).

Naturally, Lurie is not a lone voice in calling for this change of outlook. Bertrand Toën likewise gives an account of *derived algebraic geometry*:

> Derived algebraic geometry is an extension of algebraic geometry whose main purpose is to propose a setting to treat geometrically special situations (typically bad intersections, quotients by bad actions, . . .), as opposed to generic situations (transversal intersections, quotients by free and proper actions, . . .). (Toën, 2014, 1)

He explains the need for 'homotopical perturbation' in Kuhnian terms,

> the expression *homotopical mathematics* reflects a shift of paradigm in which the relation of equality relation is weakened to that of homotopy. (Toën, 2014, 3)

At the same time he points the reader to the Homotopy Type Theory and Univalent Foundation (HoTT/UF) programme as the new foundational language for this homotopical mathematics.

Now, despite this shift to what appears to be the more complex *derived* setting, familiar features are retained:

> Just as an ordinary scheme is defined to be "something which looks locally like *SpecA* where *A* is a commutative ring", a derived scheme can be described as "something which looks locally like *SpecA* where *A* is a simplicial commutative ring". (Lurie, 2009, 5)

So, an apparently complicated space is being stuck together from pieces. This theme is taken up by Carchedi in a recent paper:

> we will make precise what is means to glue structured ∞-topoi along local homeomorphisms (i.e. étale maps) starting from a collection of local models. This parallels the way one builds

manifolds out of Euclidean spaces, or schemes out of affine schemes. Since we are allowing our "spaces" to be ∞-topoi however, in these two instances we get much richer theories than just the theory of smooth manifolds, or the theory of schemes, but rather get a theory of higher generalized orbifolds and a theory of higher Deligne-Mumford stacks respectively. This same framework extends to the setting of derived and spectral geometry as well.

(Carchedi, 2013, 43)

The obvious point to be made is that all of this is just simply unthinkable without category theory. No category theory, no modern geometry of this kind. On the other hand, it may strike the reader as rather daunting that we may need to get a good handle on what Lurie, Toën, and Carchedi are doing with $(\infty, 1)$-toposes. However, we are in luck since just the right kind of foundational language is at hand to help. As Toën noted, in recent years there has emerged homotopy type theory (see Shulman's and Awodey's chapters, this volume), which is expected to play the role of the internal language of $(\infty, 1)$-toposes. Now, this language can be extended to describe large tracts of the constructions of geometry. Schreiber found the ingredients for such an extension in the writings of Lawvere, but needed to transplant them from the original topos setting to the setting of $(\infty, 1)$-toposes. Schreiber and Shulman (2014) worked out how this can be done synthetically by adding 'modalities' to homotopy type theory.

I say it is fortunate for us that this is so, but we should not underestimate the work that is still required. If we recall Friedman's scheme of meta-scientific work leading up to a revolution followed by philosophical interpretative work to make sense of it, we might say that the cycle was largely broken through the twentieth century. Even the lessons of the seventy-year-old category theory are still very far from having been absorbed within philosophy. There have been many contributions made over the decades, but not the kind of sustained work that would make it matter of course for someone entering on a career in philosophy of mathematics to know the basics of category theory. At the very least adjunctions and monads are needed to make any headway.

There will not be space to go into much detail here, but let us begin at the ordinary 1-category level with Lawvere's notion of *cohesion* (Lawvere, 2007) expressed as a chain of adjunctions between a category of spaces and the category of sets. If we take the former to be topological spaces, then one basic mapping takes such a space and gives its underlying set of points. All the cohesive 'glue' has been removed. Now there are two ways to generate a space from a set: one is to form the space with the discrete topology, where no point sticks to another; the other is to form the space with the codiscrete topology, where the points are all glued together into a single blob so that no part is separable, in the sense that there are only constant maps from a codiscrete space to the discrete space with two points. Finally, we need a second map from spaces to sets, one which 'reinforces' the glue by reducing each connected part to an element of a set, the connected components functor, π_0:

$$(\pi_0 \dashv Disc \dashv U \dashv coDisc) : Top \to Set$$

These four functors form an adjoint chain, where any of the three compositions of two adjacent functors ($U \circ coDisc$, $U \circ Disc$, $\pi_0 \circ Disc$) from the category of sets to itself is the identity, whereas, in the other direction, composing adjacent functors to produce endofunctors on *Top* ($coDisc \circ U$, $Disc \circ U$, $Disc \circ \pi_0$) yields two idempotent monads and one idempotent comonad.

Adjoint modalities where the monad is the right adjoint, $\Box \dashv \bigcirc$, can be thought of as two different opposite 'pure moments', such as codiscreteness and discreteness in this case, or in another example by Lawvere (2000), oddness and evenness of integers. There is an equivalence between types which are pure according to one of the moments and those pure according to the other, but these pure collections inject into the whole differently, as with odd and even integers into all integers, and discrete and codiscrete spaces into all spaces.

On the other hand, in adjoint modalities where the monad is on the left, $\bigcirc \dashv \Box$, there is a single moment, but the full collection of types projects onto those pure according to this moment in two different ways. Here, cohesive spaces project in two opposite ways to discrete spaces, either by the complete removal of the cohesion or by the identication of any cohering points. Another simple example has the real numbers project to the integers (entities which are purely integral) in two ways, via the floor and ceiling functions.

What Schreiber does is to find analogous modalities generated by an adjoint quadruple between an $(\infty, 1)$-topos, $\mathbf{H}$, and the base $(\infty, 1)$-topos of ∞-groupoids, $\infty Grpd$:

$$(\Pi \dashv Disc \dashv \Gamma \dashv coDisc) : \mathbf{H} \to \infty Grpd.$$

The three induced 'adjoint modalities' are called shape modality $\dashv$ flat modality $\dashv$ sharp modality and denoted $\int \dashv \flat \dashv \sharp$. In a sense, this $\mathbf{H}$ can be seen as spaces modelled on a 'thickened' point.

Now a very similar pattern repeats itself in the form of a further string of four adjunctions, this time between $\mathbf{H}$ and another $(\infty, 1)$-topos, corresponding to extending the thickened point infinitesimally. The three resulting adjoint modalities now comprise two comonads and one monad.

The existence of these two related sets of three adjoint modalities is extraordinarily powerful, allowing the expression of a rich internal higher geometry, including Galois theory, Lie theory, differential cohomology, and Chern–Weil theory, and allows for the synthetic development of higher gauge theory (Schreiber, 2013). There remains plenty more interpretative work to be done in making these ideas more accessible, but for our purposes here let us just retain the monad of the second adjoint modality triple, denoted $\Im$ or sometimes $\int_{inf}$. It is the important one for us to continue the story from Lurie and Carchedi.

So now, despite the apparently intimidating complexity of modern geometry, it is possible to maintain, as Schreiber does, that there remains a simplicity.

It would seem to me that the old intuition, seemingly falling out of use as the theory becomes more sophisticated, re-emerges strengthened within higher topos theory . . . Notably all those "generalized schemes", "étale infinity-groupoids" and so forth are nothing but the implementation of the old intuition of "big spaces glued from small model spaces" implemented in homotopy theory . . . I think it's a general pattern, in the wake of homotopy type theory we find that much of what looks super-sophisticated in modern mathematics is pretty close to the naive idea, but implemented internally in an ∞-topos. (Schreiber, 2014b)

With homotopy type theory and the six modalities briefly mentioned earlier, and in particular the infinitesimal shape modality, it is possible to describe *synthetically* what it is to be a 'formally étale morphism'. Now choosing types, $\{U_i\}$, as 'model spaces', then a general geometric space is a type X equipped with a map of the form

$$\coprod_j U_j \longrightarrow X,$$

such that this map is a 1-epimorphism and formally étale.[1] We have arrived thus at a synthetic formulation of one of the very basic ideas of geometry.

Of course, there are many such basic ideas for us to consider. In this section, I have sketched some ideas of an extraordinarily ambitious body of scientific and *meta-scientific* work. It may appear that by proposing that we understand cutting-edge geometry, I risk being caught up with the changeable fashions of research, but let us not forget that these projects are rooted in the ideas of Grothendieck from many decades ago, and that later developments were foreseen to some considerable extent by him (see, e.g., Grothendieck, 1983). Current ideas thus emerge out of a vast body of work. Indeed, Toën motivates a section where he constructs "a brief, and thus incomplete, history of the mathematical ideas that have led to the modern developments of derived algebraic geometry" as follows:

As we will see the subject has been influenced by ideas from various origins, such as intersection theory in algebraic geometry, deformation theory, abstract homotopy theory, moduli and stacks theory, stable homotopy theory, and so on. Derived algebraic geometry incorporates all these origins, and therefore possesses different facets and can be comprehended from different angles. We think that knowledge of some of the key ideas that we describe below can help to understand the subject from a philosophical as well as from technical point of view. (Toën, 2014, 6)

If some details will inevitably change, that $(\infty, 1)$-categories lie at the heart of modern geometry will very likely not.

[1] In the case of schemes, one needs to modify slightly to *pro-étale* morphisms, in some sense a reflection of the less homogeneous nature of the spaces.

2.5 Conclusion

I have sketched a broad canvas in this chapter. This is to some degree forced upon us by the state we are in where philosophy has drifted from its task. Had the course of philosophy after the famous Davos meeting (Friedman, 2000) favoured Cassirer, we might have had a generation of philosophers keen to search for the emergence of new self-understandings in mathematics. Surely in that case category theory, and its higher forms, would have been absorbed much more fully into philosophical consciousness. With the emergence of homotopy type theory, which is already generating considerable philosophical interest, we may see this happen at last. What I have described in this chapter should suggest that there is a great deal of further work to be done in coming to understand extensions of homotopy type theory, certainly the cohesive variety so far as geometry goes. It should also be noted that with a linear logic variant of homotopy type theory it is possible to express synthetically many aspects of the quantization of higher gauge theory (Schreiber, 2014a).

We have seen Weyl-like meta-scientific work in the formulation of cohesive homotopy type theory, requiring a range of modalities to be added to the basic type theory. Unlike Weyl with Fichte, Schreiber follows Lawvere (1970, 1991) in finding inspiration in Hegel. One can even tell a 'Hegelian' story starting from the opposition between $\emptyset$ and 1, rising through a process of 'Aufhebung' to the six modalities (Schreiber, 2014a, sect. 2.4), and even beyond to a further set of three modalities which may be interpreted as capturing the supergeometry needed for dealing with fermions.

Contrast this with a different kind of use of Hegel by Cassirer and also Lakatos, a philosopher more familiar to the Anglophone community. With the new framework for geometry in place, we should be able to tell the Cassirerian story of the unfolding of the past in mathematics and physics, as mathematicians such as Toën are inclined to do by themselves. Mathematics is to be understood by the fact that it constitutes a single tradition of intellectual enquiry. Ideas found at particular stages possess the seeds of later formulations, which retrospectively allow us to understand them better.

We can use this opportunity to gain a grip on some real mathematical content, offering the opportunity for a more interesting dialogue between philosophy of physics and philosophy of mathematics. For one thing, the duality between geometry and algebra that we saw between rings and affine schemes, and which lies behind the relation between the Heisenberg and Schrödinger pictures, continues to higher geometry and higher algebra, where it manifests itself in different formulations of higher gauge theory. Fundamentally, this duality relates to the operation of taking opposities of $(\infty, 1)$-categories (Corfield, 2015).

Finally, as with Cassirer's observation about the seeds of the Erlanger Programme lying within our perception, it is sometimes revealed during and after moments of synthesis in mathematics that there is a reliance on aspects of cognition, perception, and language, which had possibly gone unnoticed. I think at the very least a form of dependent type theory is present in our cognition as manifested in ordinary language

(Ranta, 1995). Likewise, the idea of big spaces glued from small model spaces seems very basic. It is surely no accident that mathematicians speak of an 'atlas' to define a manifold, since an ordinary atlas provides a collection of maps which overlap. It seems likely we employ something like this in the cognitive maps by which we navigate our domain. Perhaps one of the invariants of geometry has been found here.

References

Atiyah, M. (2003). What is geometry? in *The Changing Shape of Geometry: Celebrating a Century of Geometry and Geometry Teaching*, 24–30. Cambridge University Press, Cambridge, UK.

Ben-Zvi, D. (2014). Blog comment. https://www.math.columbia.edu/ woit/wordpress/?p=7114&cpage=1# comment-214353

Borger, J. (2009). Blog comment. https://golem.ph.utexas.edu/category/2009/02/lakatos_as_dialectical_realist.html#c022225

Carchedi, D. (2013). Higher orbifolds and Deligne-Mumford stacks as structured infinity topoi. Arxiv preprint: http://arxiv.org/abs/1312.2204

Cassirer, E. (1944). The concept of group and the theory of perception. *Philosophy and Phenomenological Research* 5(1), 1–36.

Cassirer, E. (1957). *The Philosophy of Symbolic Forms: The Phenomenology of Knowledge*, Vol. 3. Yale University Press, New Haven, CT.

Corfield, D. (2003). *Towards a Philosophy of Real Mathematics*. Cambridge University Press, Cambridge, UK.

Corfield, D. (2012). Narrative and the rationality of mathematical practice, in A. Doxiadis and B. Mazur (eds) *Circles Disturbed: The Interplay of Mathematics and Narrative*, 244–80. Princeton University Press, Princeton, NJ.

Corfield, D. (2015). Duality as a category-theoretic concept. *Studies in History and Philosophy of Modern Physics*, doi:10.1016/j.shpsb.2015.07.004.

Domski, M., and Dickson, M. (eds) (2010). *Discourse on a New Method: Reinvigorating the Marriage of History and Philosophy of Science*. Open Court Publishing Company.

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of Science* 1(2), 163–9.

Friedman, M. (2000). *The Parting of the Ways: Carnap, Cassirer, and Heidegger*. Open Court Publishing Company.

Friedman, M. (2001). *The Dynamics of Reason*. University of Chicago Press, Chicago.

Giovanelli, M. (2013). Talking at cross-purposes: how Einstein and the logical empiricists never agreed on what they were disagreeing about. *Synthese* 190(17), 3819–63.

Grothendieck, A. (1983). Pursuing stacks, Letter to D. Quillen, in G. Maltsiniotis, M. Künzer, and B. Toen (eds), *Documents Mathématiques*. Society Mathematics Paris, France.

Grothendieck, A., and Verdier, J. (1972). *Théorie des topos et cohomologie étale des schémas*, Tome 1: Théorie des topos. *Lecture Notes in Mathematics*, Vol. 269. Springer-Verlag, Berlin.

Heis, J. (2011). Ernst Cassirer's neo-Kantian philosophy of geometry. *British Journal for the History of Philosophy* 19(4), 759–94.

Lawvere, W. (1970). Quantifiers and sheaves. *Actes du congres international des mathematiciens, Nice* 1, 329–34.

Lawvere, W. (1991). Some thoughts on the future of category theory, in A. Carboni et. al. (eds), *Category Theory*, 1–13 Springer, New York.

Lawvere, W. (2000). Adjoint cylinders. Categories mailing list comment, http://permalink.gmane.org/gmane.science.mathematics.categories/1683

Lawvere, W. (2007). Axiomatic cohesion. *Theory and Applications of Categories* 19(3), 41–9.

Lurie, J. (2009). Derived algebraic geometry V: structured spaces. Arxiv preprint: http://arxiv.org/abs/0905.0459

Marquis, J.-P. (2008). *From a Geometric Point of View: A Study of the History and Philosophy of Category Theory*. Springer, New York.

McLarty, C. (2008). There is no ontology here: visual and structural geometry in arithmetic, in P. Mancosu (ed.), *The Philosophy of Mathematical Practice*, 370–406. Oxford University Press, Oxford.

MRSI (2014). Geometric Representation Theory, Programme announcement, https://www.msri.org/programs/276

Nagel, E. (1939). The formation of modern conceptions of formal logic in the development of geometry. *Osiris* 7, 142–223.

Ranta, A. (1995). *Type-Theoretical Grammar*. Oxford University Press, Oxford.

Ryckman, T. (2005). *The Reign of Relativity*. Oxford University Press, Oxford.

Scholz, E. (2005). Philosophy as a cultural resource and medium of reflection for Hermann Weyl. *Revue de Synthèse* 126, 331–351.

Scholz, E. (2011). H. Weyl's and E. Cartan's proposals for infinitesimal geometry in the early 1920s. *Boletim da Sociedada Portuguesa de Matematica* Numero Especial A, 225–45.

Schreiber, U. (2013). Differential cohomology in a cohesive infinity-topos. Arxiv preprint: http://arxiv.org/abs/1310.7930

Schreiber, U. (2014a). Quantization via Linear homotopy types. Arxiv preprint: http://arxiv.org/abs/1402.7041

Schreiber, U. (2014b). nForum discussion comment. http://nforum.ncatlab.org/discussion/2084/higher-geometry/?Focus=49931# Comment_49931

Schreiber, U., and Shulman, M. (2014). Quantum gauge field theory in cohesive homotopy type theory. Arxiv preprint: http://arxiv.org/abs/1408.0054

Toën, B. (2014). Derived algebraic geometry. Arxiv preprint: http://arxiv.org/abs/1401.1044

Torretti, R. (1978). *Philosophy of Geometry from Riemann to Poincaré*. Springer, New York.

3

Homotopy Type Theory: A Synthetic Approach to Higher Equalities

Michael Shulman

3.1 Introduction

Ask an average mathematician or philosopher today about the foundations of mathematics, and you are likely to receive an answer involving set theory: an apparent consensus in marked contrast to the foundational debates of the early twentieth century. Now, at the turn of the twenty-first century, a new theory has emerged to challenge the foundational ascendancy of sets. Arising from a surprising synthesis of constructive intensional type theory and abstract homotopy theory, Homotopy Type Theory and Univalent Foundations (HoTT/UF) purports to represent more faithfully the everyday practice of mathematics, but also provides powerful new tools and a new paradigm. So far, its concrete influence has been small, but its potential implications for mathematics and philosophy are profound.

There are many different aspects to HoTT/UF,[1] but in this chapter I will focus on its use as a foundation for mathematics. Like set theory, it proposes to found mathematics on a notion of *collection*, but its collections (called *types*) behave somewhat differently. The most important difference is that in addition to having elements as sets do, the types of HoTT/UF come with further collections of *identifications* between these elements (i.e. ways or reasons that they are equal). These identifications form a structure that modern mathematicians call an ∞-*groupoid* or *homotopy type*, which is a basic object of study in homotopy theory and higher category theory; thus, HoTT/UF offers mathematicians a new approach to the latter subjects.

Of greater importance philosophically, however, is HoTT/UF's proposal that such types can be the fundamental objects out of which mathematics and logic are built.

[1] Though HoTT and UF are not identical, the researchers working on both form a single community, and the boundary between them is fluid. Thus, I will not attempt to distinguish between them, even if it results in some technically incorrect statements.

In other words, HoTT/UF suggests that whenever we mentally form a collection of things, we must *simultaneously* entertain a notion of what it means for two of those things to be the same (in contrast to the position of Zermelo-Fraenkel theory with choice (ZFC) that all things have an identity criterion *prior* to their being collected into a set). As stated, this is closely related to the conception of "set" promulgated by Bishop, but HoTT/UF generalizes it by allowing two things to "be the same" in *more than one way*. This is perhaps not a common everyday occurrence, but it is a fundamental part of category theory and thus an integral part of mathematics, including many modern theories of physics. Thus, like other initially unintuitive ideas such as relativistic time dilation and quantum entanglement, it can be argued to be basic to the nature of reality. The innovation of HoTT/UF is that this idea can be made basic to the foundational logical structure of mathematics as well, and that doing so actually *simplifies* the theory.

In this chapter, I will attempt to convey some of the flavor and advantages of HoTT/UF; we will see that in addition to expanding the discourse of mathematics, it also represents certain aspects of *current* mathematical practice more faithfully than set theory does. In Sections 3.2 and 3.3, I will describe HoTT/UF very informally; in Sections 3.4–3.6, I will discuss some of its features in a bit more detail; and in Section 3.7, I will attempt to pull together all the threads with an example. For space reasons, I will not be very precise, nor will I discuss the history of the subject in any depth; for more details, see Univalent Foundations Program (2013). Other recent survey articles on HoTT/UF include Awodey (2012); Awodey et al. (2013); and Pelayo and Warren (2014).

For helpful conversations and feedback, I would like to thank (in random order) Emily Riehl, David Corfield, Dimitris Tsementzis, James Ladyman, Richard Williamson, Martín Escardó, Andrei Rodin, Urs Schreiber, John Baez, and Steve Awodey, as well as numerous other contributors at the *n*-Category Café and the HoTT email list, and the referees.

3.2 ∞-groupoids

The word "∞-groupoid" looks complicated, but the underlying idea is extremely simple, arising naturally from a careful consideration of what it means for two things to be "the same". Specifically, it happens frequently in mathematics that we want to define a collection of objects that are determined by some kind of "presentation", but where "the same" object may have more than one presentation. As a simple example, if we try to define a *real number* to be an infinite decimal expansion[2] such as $\pi = 3.14159\cdots$, we encounter the problem that (for instance)

[2] Like any mathematical object, there are many equivalent ways to define the real numbers. This specific definition is rarely used in mathematics for technical reasons, but it serves as a good illustration, and the common definition of real numbers using Cauchy sequences has exactly the same issues.

$$0.5 = 0.50000\cdots \qquad \text{and} \qquad 0.4\overline{9} = 0.49999\cdots$$

are distinct decimal expansions but ought to represent the same real number. Therefore, "the collection of infinite decimal expansions" is not a correct way to define "the collection of real numbers".

If by "collection" we mean "set" in the sense of ZFC, then we can handle this by defining a real number to be a *set* of decimal expansions that all "define the same number", and which is "maximal" in that there are no *other* expansions that define the same number. Thus, one such set is $\{0.5, 0.4\overline{9}\}$, and another is $\{0.\overline{3}\}$. These sets are *equivalence classes*, and the information about which expansions define the same number is an *equivalence relation* (a binary relation $\sim$ such that $x \sim x$, if $x \sim y$ then $y \sim x$, and if $x \sim y$ and $y \sim z$ then $x \sim z$). The set of equivalence classes is the *quotient* of the equivalence relation.

Similarly, Frege (1884, sect. 68) defined the *cardinality* of a set X to be (roughly, in modern language) the set of all sets related to X by a bijection. Thus, for instance, 0 is the set of all sets with no elements, 1 is the set of all singleton sets, and so on. These are exactly the equivalence classes for the equivalence relation of bijectiveness. That is, we consider a cardinal number to be "presented" by a set having that cardinality, with two sets presenting the same cardinal number just when they are bijective.

An example outside of pure mathematics involves Einstein's theory of general relativity, in which the universe is represented by a differentiable manifold with a metric structure. In this theory, if two manifolds are *isomorphic* respecting their metric structure, then they represent the same physical reality. (An isomorphism of manifolds is often called a "diffeomorphism", and if it respects the metric it is called an "isometry".) Thus we find, for instance, in Sachs and Wu (1977, sect. 1.3) that

> A general relativistic *gravitational field* $[(M, \mathbf{g})]$ is an equivalence class of spacetimes [manifolds M with metrics $\mathbf{g}$] where the equivalence is defined by . . . isometries.

This sort of situation, where multiple mathematical objects represent the same physical reality, is common in modern physics, and the mathematical objects (here, the manifolds) are often called *gauges*.[3]

Definitions by equivalence classes are thus very common in mathematics and its applications, but they are not the only game in town. A different approach to the problem of "presentations" was proposed by Bishop (1967, sect. 1.1):

> A set is defined by describing exactly what must be done in order to construct an element of the set and what must be done in order to show that two elements are equal.

In other words, according to Bishop, a *set* is a collection of things *together with* the information of when two of those things are equal (which must be an equivalence

[3] Whether general relativity should be technically considered a "gauge theory" is a matter of some debate, but all that matters for us is that it exhibits the same general phenomenon of multiple models.

relation).[4] Thus, the real numbers would *be* infinite decimal expansions, but "the set of real numbers" would include the information that (for instance) 0.5 and $0.4\overline{9}$ are the same real number. One advantage of this is that if we are given "a real number", we never need to worry about *choosing* a decimal expansion to represent it. (Of course, for decimal expansions there are canonical ways to make such a choice, but in other examples there are not.)

As a much older example of this style of definition, in Euclid's *Elements* we find the following:

> **Definition 4.** Magnitudes are said to *have a ratio* to one another which can, when multiplied, exceed one another.
>
> **Definition 5.** Magnitudes are said to be *in the same ratio*, the first to the second and the third to the fourth, when, if any equimultiples whatever are taken of the first and third, and any equimultiples whatever of the second and fourth, the former equimultiples alike exceed, are alike equal to, or alike fall short of, the latter equimultiples respectively taken in corresponding order.

That is, Euclid first defined how to *construct* a ratio, and then second he defined when two ratios are *equal*, exactly as Bishop says he ought.

On its own, Bishop's conception of set is not a very radical change. But it paves the way for our crucial next step, which is to recognize that frequently there may be more than one "reason" why two "presentations" define the same object. For example, there are two bijections between $\{a, b\}$ and $\{c, d\}$: one that sends a to c and b to d, and another that sends a to d and b to c. Likewise, a pair of manifolds may be isometric in more than one way.

This should not be confused with the question of whether there is more than one *proof* that two things are the same. Rather, the question is whether substituting one for the other in a mathematical statement or construction can yield multiple inequivalent results. For instance, there is a predicate P on $\{a, b\}$ such that $P(a)$ is true and $P(b)$ is false. We can "transport" P along a bijection from $\{a, b\}$ to $\{c, d\}$ to obtain a predicate Q on $\{c, d\}$, but the resulting Q will depend on which bijection we use. If we use the bijection that sends a to c and b to d, then $Q(c)$ will be true and $Q(d)$ will be false, but if we use the other bijection, then $Q(c)$ will be false and $Q(d)$ will be true. Thus, $\{a, b\}$ and $\{c, d\}$ "are the same" in more than one way.

If a predicate or construction is left literally unchanged by this sort of substitution, it is called *invariant*. Thus, physicists speak of *gauge invariance* when talking about theories with multiple mathematical models of the same reality. More generally, a construction that "varies appropriately" under such substitutions (but in a way potentially dependent on the "reason" for sameness, as explained earlier) is called

[4] Although Bishop's goal was to give a constructive treatment of mathematics, this notion of "set" is meaningful independently of whether one's logic is constructive or classical.

covariant. In particular, general relativity is said to be *generally covariant*, meaning that a mathematical model of reality can be replaced by any isometric one—but in a way dependent on the particular isometry chosen.

This behavior lies at the root of Einstein's famous *hole argument*, which can be explained most clearly as follows. Suppose M and N are manifolds with spacetime metrics $\mathbf{g}$ and $\mathbf{h}$, respectively, and ϕ is an isometry between them. Then any point $x \in M$ corresponds to a unique point $\phi(x) \in N$, both of which represent the same "event" in spacetime. Since ϕ is an isometry, the gravitational field around x in M is identical to that around $\phi(x)$ in N. However, if ψ is a *different* isomorphism from M to N which does *not* respect the metrics, then the gravitational field around x in M may be quite different from that around $\psi(x)$ in N.

So far, this should seem fairly obvious. But Einstein originally considered only the special case where M and N happened to be the same manifold (though not with the same metric), where ψ was the identity map id_M, and where ϕ was the identity outside of a small "hole". In this case, it seemed wrong that two metrics could be the same outside the hole but different inside of it. The solution is clear from the more general situation in the previous paragraph: the fact that the two metrics "represent the same reality" is witnessed by the isomorphism ϕ, not ψ. Thus, even for a point x inside the hole, we should be comparing $\mathbf{g}$ at x with $\mathbf{h}$ at $\phi(x)$, not with $\mathbf{h}$ at $\mathrm{id}_M(x) = x$.[5]

This and other examples show that it is often essential to *remember which* isomorphism we are using to treat two objects as the same. The set-theoretic notion of equivalence classes is unable to do this, but Bishop's approach can be generalized to handle it. Indeed, such a generalization is arguably already latent in Bishop's constructive phrasing: both the construction of elements and the proofs of equality are described in terms of *what must be done*, so it seems evident that just as there may be more than one way to construct an element of a set, there may be more than one way to show that two elements are equal. Bishop made no use of this possibility, but HoTT/UF takes it seriously. The laws of an equivalence relation then become algebraic structure on these "reasons for equality": given a way in which $x = y$ and a way in which $y = z$, we must have an induced way in which $x = z$, and so on, satisfying natural axioms. The resulting structure is called a *groupoid*. Thus, for instance, spacetime manifolds form a groupoid, in which the ways that $M = N$ are the isometries from M to N (if any exist).

If it should happen that for every x and y in some groupoid, there is *at most one* reason why $x = y$, then our groupoid is essentially just a set in Bishop's sense; thus, the universe of sets is properly included in that of groupoids. This is what happens with decimal expansions: there is only one way in which 0.5 and $0.4\overline{9}$ represent the same real number (i.e. in any statement or construction involving 0.5, there is only one way to replace 0.5 by $0.4\overline{9}$). This is in contrast to the situation with manifolds, where using

[5] While this description in modern language makes it clear why there is no paradox, it does obscure the reasons why for many years people *thought* there was a paradox! I will return to this in Section 3.7.

a different isomorphism ϕ or ψ from M to N can result in different statements, e.g. one which speaks about $\phi(x) \in N$ and another about $\psi(x) \in N$.

The final step of generalization is to notice that we introduced sets (and generalized them to groupoids) to formalize the idea of "collection", but we have now introduced, for each pair of things x and y in a groupoid, an *additional* collection, namely the ways in which x and y are equal. Thus, it seems natural that this collection should itself be a set, or more generally a groupoid, so that two ways in which $x = y$ could themselves be equal or not, and perhaps in more than one way. Taken to its logical conclusion, this observation demands an infinite tower consisting of elements, ways in which they are equal, ways in which those are equal, ways in which *those* are equal, and so on. Together with all the necessary operations that generalize the laws of an equivalence relation, this structure is what we call an *∞-groupoid*.

This notion may seem very abstruse, but over the past few decades ∞-groupoids have risen to a central role in mathematics and even physics, starting from algebraic topology and metastasizing outwards into commutative algebra, algebraic geometry, differential geometry, gauge field theory, computer science, logic, and even combinatorics. It turns out to be very common that two things can be equal in more than one way.

3.3 Foundations for Mathematics

In Section 3.2, I introduced the notion of ∞-groupoid informally. At this point a modern mathematician would probably try to give a *definition* of ∞-groupoid, such as "an ∞-groupoid consists of a collection of elements, together with for any two elements x, y a collection of ways in which $x = y$, and for any two such ways f, g a collection of ways in which $f = g$, and so on, plus operations …". Clearly, any such definition must refer to a *prior* notion of "collection", which a modern mathematician would probably interpret as "set". Such definitions of ∞-groupoids are commonly used, although they are quite combinatorially complicated.

However, in Section 3.2, we considered ∞-groupoids not as *defined in terms of* sets, but as *substitutes* or rather *generalizations* of them. Thus, we should instead seek a theory at roughly the same ontological level as ZFC, whose basic objects are ∞-groupoids. This is exactly what HoTT/UF is: a *synthetic theory of ∞-groupoids*.[6]

The word "synthetic" here is, as usual, used in opposition to "analytic". In modern mathematics, an analytic theory is one whose basic objects are defined in some other theory, whereas a synthetic theory is one whose basic objects are undefined terms given meaning by rules and axioms. For example, *analytic geometry* defines points

[6] Since ∞-groupoids are a formalization of the idea that things can be equal in more than one way, that these ways can themselves be equal in more than one way, and so on, we may equivalently (but more informally) call HoTT/UF a *synthetic theory of higher equalities*, as in the chapter title.

and lines in terms of numbers, whereas *synthetic geometry* is like Euclid's with "point" and "line" essentially undefined.[7]

Thus, our first step to understanding HoTT/UF is that it is an axiomatic system in which "∞-groupoid" is essentially an undefined term. One advantage of this can already be appreciated: it allows us to say simply that for any two elements x and y of an ∞-groupoid, the "ways in which $x = y$" form another ∞-groupoid, so that ∞-groupoids are really the only notion of "collection" that we need consider. As part of a *definition* of ∞-groupoid, this would appear circular, but as an *axiom*, it is unobjectionable.

So far, this description of HoTT/UF could also be applied (with different terminology) to the field of mathematics called "abstract homotopy theory". However, although HoTT/UF is strongly influenced by homotopy theory, there is more to it: as suggested earlier, its ∞-groupoids can substitute for sets as a foundation for mathematics.

When I say that a synthetic theory can be a *foundation for mathematics*, I mean simply that we can encode the rest of mathematics into it somehow.[8] This definition of "foundation" is reasonably precise and objective, and agrees with its common usage by most mathematicians. A computer scientist might describe such a theory as "mathematics-complete", by analogy with Turing-complete programming languages (that can simulate all other languages) and NP-complete problems (that can solve all other NP problems). For example, it is commonly accepted that ZFC set theory has this property. On the other hand, category theory in its role as an organizing principle for mathematics, though of undoubted philosophical interest, is not foundational in this sense (although a synthetic form of category theory like that of Lawvere (1966) could be).

In particular, a synthetic theory cannot fail to be foundational because some analytic theory describes similar objects. The fact that we *can* define and study ∞-groupoids inside of set theory says nothing about whether a *synthetic* theory of ∞-groupoids can be foundational. To the contrary, in fact, it is highly *desirable* of a new foundational theory that we can translate back and forth to previously existing foundations; among other things it ensures the relative consistency of the new theory. Similarly, we cannot dismiss a new synthetic foundational theory by claiming that it "requires some pre-existing notions": the simple fact of being synthetic means that it does not. Of course, humans always try first to *understand* new ideas in terms of old ones, but that doesn't make the new ideas *intrinsically* dependent on the old. A student may learn that dinosaurs are like "big lizards", but that doesn't make lizards logically, historically, or genetically prior to dinosaurs.

[7] Euclid's *Elements* as they have come down to us do contain "definitions" of "point" and "line", but these are not definitions in a modern mathematical sense, and more modern versions of Euclidean geometry such as that of Hilbert (1899) do leave these words undefined.

[8] Or into some natural variant or extension of it, such as by making the logic intuitionistic or adding stronger axioms.

In addition, we should beware of judging a theory to be more intuitive or fundamental merely because we are familiar with it: intuition is not fixed, but can be (and is) trained and developed. At present, most mathematicians think of ∞-groupoids in terms of sets because they learned about sets early in their mathematical education; but even in its short existence the HoTT/UF community has already observed that graduate students who are "brought up" thinking in HoTT/UF form a direct understanding and intuition for it that sometimes outstrips that of those who "came to it late". Moreover, the ZFC-like intuitions about set theory now possessed by most mathematicians and philosophers also had to be developed over time: Lawvere (1994) has pointed out that Cantor's original "sets" seem more like those of Lawvere's alternative set theory, the Elementary Theory of the Category of Sets (ETCS) (see Lawvere (2005) and McLarty's chapter in the present volume).

The point being made, therefore, is that HoTT/UF, the synthetic theory of ∞-groupoids, can be a foundation for mathematics in this sense. There is quite an easy proof of this: we have already seen that the universe of ∞-groupoids properly contains a universe of sets. More precisely, there is a subclass of the ∞-groupoids of HoTT/UF which together satisfy the axioms of ETCS.[9] A model of ZFC can then be constructed using trees as described in McLarty's chapter, or directly as in Univalent Foundations Program (2013, sect. 10.5). Thus, any mathematics that can be encoded into set theory can also be encoded into HoTT/UF. (Of course, if we intended to encode *all* of mathematics into HoTT/UF via set theory this way, there would be no benefit to choosing HoTT/UF as a foundation over set theory. The point is that *some* parts of mathematics can be also encoded into HoTT/UF in *other*, perhaps more natural, ways.)

In sum, if we so desire, *we may regard the basic objects of mathematics to be ∞-groupoids rather than sets*. Our discussion in Section 3.2 suggests some reasons why we might want to do this; I will mention some further advantages as they arise. But it is now time to say something about what HoTT/UF actually looks like.

3.4 Type Theory and Logic

The basic objects of HoTT/UF behave like ∞-groupoids; but we generally call them *types* instead, and from now on I will switch to this usage. This particular word is due to the theory's origins in Martin-Löf type theory (Martin-Löf, 1975); but (in addition to being five syllables shorter) it also fortuitously evokes the terminology "homotopy type" from algebraic topology, which is essentially another word for "∞-groupoid" (see e.g. Baez, 2007).

[9] In fact, HoTT/UF is not (yet) a single precisely specified theory like ZFC and ETCS: as befits a young field, there are many variant theories in use and new ones under development. In particular, when I say "HoTT/UF" I mean to encompass both "classical" versions that have the Axiom of Choice and Law of Excluded Middle and also "intuitionistic" or "constructive" ones that do not. In the latter cases, the universe of sets satisfies not ETCS (which is classical) but an "intuitionistic" version thereof.

Like sets, the types of HoTT/UF have *elements*, also called *points*. We write $x : A$ when x is a point of A; the most salient difference between this and ZFC's "$x \in A$" is that (like in ETCS) we cannot compare elements of different types: a point is always *a point of some type*, that type being part of its nature. Whenever we introduce a variable, we must specify its type: whereas in ZFC "for every integer x, $x^2 \geq 0$" is shorthand for "for every thing x, if x happens to be an integer then $x^2 \geq 0$", in HoTT/UF the phrase "for every integer x" is atomic. This arguably matches mathematical practice more closely, although the difference is small.

The basic theory of HoTT/UF is a collection of *rules* stipulating operations we can perform on types and their points. For instance, if A and B are types, there is another type called their cartesian product and denoted $A \times B$. Any such rule for making new types comes with some number of rules for making points of these types: in the case of products, this rule is that given $a : A$ and $b : B$, we have an induced point of $A \times B$ denoted (a, b). We also have dual rules for extracting information from points of types, e.g. from any $x : A \times B$ we can extract $\pi_1(x) : A$ and $\pi_2(x) : B$. Of course, $\pi_1(a, b)$ is a and $\pi_2(a, b)$ is b.

It is important to understand that these *rules* are not the same sort of thing as the *axioms* of a theory like ZFC or ETCS. Axioms are statements *inside* an ambient superstructure of (usually first-order) logic, whereas the rules of type theory exist at the same level as the deductive system of the logic itself. In a logic-based theory like ZFC, the "basic act of mathematics" is to deduce a conclusion from known facts using one of the rules of logic, with axioms providing the initial "known facts" to get started. By contrast, in a type theory like HoTT/UF, the "basic acts of mathematics" are specified directly by the rules of the theory, such as the rule for cartesian products which permits us to construct (x, y) once we have x and y. Put differently, choosing the axioms of ZFC is like choosing the starting position of a board game whose rules are known in advance, whereas choosing the rules of HoTT/UF is like choosing the rules of the game itself.

To understand the effect this distinction has on mathematical practice, we observe that the everyday practice of mathematics can already be separated into two basic activities: constructing (a.k.a. defining or specifying) and proving. For instance, an analyst may first construct a particular function, then prove that it is continuous. This distinction can be found as far back as Euclid, whose Postulates and Propositions are phrased as things to be *done* ("to draw a circle with any center and radius") rather than statements of existence, and which are "demonstrated" by making a *construction* and then *proving* that it has the desired properties. Rodin (2017) has recently argued that this distinction is closely related to Hilbert's contrast between *genetic* and *axiomatic* methods.[10]

[10] At least in Hilbert and Bernays (1934–1939); in Hilbert (1900) the same words seem to refer instead to analytic and synthetic theories, respectively.

When encoding mathematics into ZFC, however, the "construction" aspect of mathematics gets short shrift, because in fully formal ZFC the only thing we *can* do is prove theorems. Thus, the encoding process must translate constructions into proofs of existence. By contrast, in HoTT/UF and other type theories like it, it appears that the pendulum has swung the other way: the *only* thing we can do is perform constructions. How, then, do we encode proofs?

The answer begins with an idea called *propositions as types*: we interpret every *statement* that we might want to prove as a *type*, in such a way that it makes sense to interpret *constructing an element* of that type as *proving the original statement*. In this way we obtain a form of logic *inside of* type theory, rather than starting with a background logic as is done in set theory. Thus, as a foundation for mathematics, type theory is "closer to the bottom" than set theory: rather than building on the same "sub-foundations" (first-order logic), we "re-excavate" the sub-foundations and incorporate them into the foundational theory itself. In the words of Pieter Hofstra, type theory is "the engine and the fuel all in one".

One reason this idea is so useful is an observation called the *Curry–Howard correspondence* (Curry, 1934; Martin-Löf, 1975; Howard, 1980; Wadler, 2015): the logical connectives and quantifiers are *already present* in type theory as constructions on types. For instance, if A and B are types representing propositions P and Q, respectively, then $A \times B$ represents the conjunction $P \wedge Q$. This is justified because the way we construct an element of $A \times B$—by constructing an element of A and an element of B—corresponds precisely to the way we prove $P \wedge Q$—by proving P and also proving Q. Similarly, the type of functions from A to B (usually denoted $A \to B$) represents the implication $P \to Q$, and so on.

If we interpret logic directly according to this correspondence, we find that just as with the encoding into ZFC, the distinction between construction and proof is destroyed; only this time it is because we must encode proofs as constructions rather than vice versa. Whereas in ZFC we cannot construct objects, only prove that they exist, under Curry–Howard we cannot prove that something exists without constructing it.

The innovation of HoTT/UF is to allow both kinds of existence to coexist smoothly. We follow the overall philosophy of propositions-as-types, but in addition we single out a small but important class of types: those that have at most one point, with no higher equality information.[11] I will call these types *truth values*, since we think of them as representing "false" (if empty) or "true" (if inhabited); they are also often called *propositions* or *mere propositions*. Moreover, we add a rule that for any type A there is a *truncation* $\|A\|$ (also called the *bracket* or *squash*), such that $\|A\|$ is a truth value, and such that given any $a : A$ we have $|a| : \|A\|$. (Since $\|A\|$ is a truth value, $|a|$ doesn't depend on the value of a, only that we have it.)

[11] The importance of these types has been particularly advocated by Voevodsky, building on precursors such as Constable et al. (1986) and Awodey and Bauer (2004).

Now we can distinguish between existence proofs and constructions by whether the type of the result is truncated or not. When we construct an element of a type A that is not a truth value, we are defining some specific object; but if we instead construct an element of $\|A\|$, we are "proving" that some element of A exists without specifying it.[12] From this point of view, which is shared by many members of the HoTT/UF community, it is misleading to think of propositions-as-types as "encoding first-order logic in type theory". While this description can serve as a first approximation, it leads one to ask and argue about questions like "should the statement $\exists x : A$ be encoded by the type A or the type $\|A\|$?" We regard this question as invalid, because it implicitly assumes that mathematics has already been encoded into first-order logic, with constructions and pure-existence proofs collapsed into the quantifier $\exists$. We reject this assumption: the proper approach is to encode *mathematics* directly into HoTT/UF, representing a construction of an element of A by the type A itself, and a pure-existence statement by its truncation $\|A\|$.

It is true that due to the ascendancy of ZFC and first-order logic in general, most modern mathematicians "think in first-order logic" and are not used to distinguishing constructions from existence proofs. However, it remains true that some kinds of theorem, such as "A is isomorphic to B", are almost always "proven" by giving a construction, and a careful analysis reveals that such "proofs" must convey more information than mere existence, because frequently one needs to know later on exactly *what* isomorphism was constructed. This is one of the ways in which HoTT/UF represents the actual practice of mathematics more faithfully than other contenders. With a little bit of practice, and careful use of language, we can learn to consciously use this feature when doing mathematics based on HoTT/UF.

By the way, while the distinction between construction and proof is sometimes identified with the opposition between constructive/intuitionistic and classical logic (as is suggested by the shared root "construct"), the relationship between the two is actually limited. On one hand, while it is true that the "natural" logic obtained by Curry–Howard turns out to be intuitionistic, one can add additional axioms that are not "constructive" but can nevertheless be used in "constructions". Indeed, the exceedingly nonconstructive Axiom of Choice asserts exactly that objects which merely exist can nevertheless be assumed to be specified, i.e. "constructed" in a formal sense. In particular, axioms of classical logic can consistently be included in HoTT/UF.

On the other hand, intuitionistic first-order logic includes "pure unspecified existence" just like classical logic does, and constructive/intuitionistic set theory (Beeson, 1985; Aczel and Rathjen, 2000/1) collapses constructions into proofs just like ZFC does. It is true that constructive mathematicians in the tradition of Martin-Löf (1975) do adhere intentionally to the original Curry–Howard interpretation, regarding it as part of their constructivism; but they must also separately refrain from using any

[12] The possibility of these two interpretations of existence was actually already noticed by Howard (1980, sect. 12).