Homotopy Type Theory:
A synthetic approach to higher equalities

Michael Shulman*

Contents

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 20th century. Now, at the turn of the 21st 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

[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.

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. 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 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 §§2–3 I will describe HoTT/UF very informally; in §§4–6 I will discuss some of its features in a bit more detail; and in §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 [36]. Other recent survey articles on HoTT/UF include [3, 2, 31].

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.

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)

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

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 [13, §68] defined the *cardinality* of a set X to be (roughly, in modern language) the set of all sets that are 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.

---

[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.

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 [33, §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 [9, §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 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:

> **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 secondly 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 above), is called *covariant*. In particular,

---

[3]Whether or not 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.

[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.

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 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.

---

[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 §7.

3 Foundations for mathematics

In §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 §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 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 above, 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 [20] 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.

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

---

[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.

[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 Hilbert's [15] 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.

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 [23] has pointed out that Cantor's original "sets" seem more like those of Lawvere's alternative set theory ETCS (see [22] 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 [36, §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 §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.

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 [28]; 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. [6]).

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 one or more 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.

---

[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 satisfy not ETCS (which is classical) but an "intuitionistic" version thereof.

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 [32] has recently argued that this distinction is closely related to Hilbert's contrast between *genetic* and *axiomatic* methods.[10]

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* [11, 18, 28, 38]: 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 have to 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.)

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

---

[10]At least in [14]; in [16] the same words seem to refer instead to analytic and synthetic theories respectively.

[11]The importance of these types has been particularly advocated by Voevodsky, building on precursors such as [10, 5].

[12]The possibility of these two interpretations of existence was actually already noticed by Howard [18, §12].

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" have to 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 [7, 1] collapses constructions into proofs just like ZFC does. It is true that constructive mathematicians in the tradition of Martin-Löf [28] 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 nonconstructive principles. That is, a constructive *philosophy* may lead one to prefer "constructions" to proofs, but this is a separate thing from the (intuitionistic) *logic* that it also leads one to prefer. Moreover, Escardó has recently argued that Brouwer himself must have intended some notion of unspecified existence, since his famous theorem that all functions $\mathbb{N}^{\mathbb{N}} \to \mathbb{N}$ are continuous is actually *inconsistent* under unmodified Curry–Howard [29].

A last aspect of type theory that deserves mention is its computational character: its rules can also be read as defining a programming language that can actually be executed by a computer. This makes it especially convenient for computer formalization of mathematical proofs, as well as for mathematical verification of computer software. Thus, HOTT/UF is also better-adapted to these purposes than set theory is,[13] and indeed computer formalization has been very significant in the origins and development of HOTT/UF. But this would fill a whole chapter by itself, so reluctantly I will say no more about it here.

5   Identifications and equivalences

So far, I have not really said anything that is unique to HOTT/UF. The description of types, rules, and elements in §4 applies to any type theory, including Martin-Löf's original one. The approach to logic using truncations is more novel, but it still does not depend on regarding types as ∞-groupoids. However, this kind of logic is particularly appropriate in HOTT/UF, for several reasons.

The first is that, like our considerations in §2, it drives us inexorably from sets to ∞-groupoids. Namely, if statements are interpreted by types, then in particular for any $x : A$ and $y : A$, the statement "$x = y$" must be *a type*, whose points we refer to as *identifications* of x with y. If A is a set, then this type is a mere truth value, but in general there is no reason for it to be so.

Somewhat magically, it turns out that the "most natural" rule governing the type $x = y$, as first given by Martin-Löf [28], does *not* imply that it is always a truth value, but *does* imply that it automatically inherits the structure of an ∞-groupoid [26, 37]. This rule is related to Leibniz's "indiscernibility of identicals", but its form is rather that of Lawvere [21], who characterized equality using an adjunction between unary predicates and binary relations. Martin-Löf's version says that if we have a type family $C(x, y, p)$ depending on x, y, and an identification $p : x = y$, and if we want to construct an element of $C(x, y, p)$ for every x, y, and p, then it suffices to construct elements of $C(x, x, \mathsf{refl}_x)$ for every x. (In particular, the proofs of transitivity and

---

[13]Although finding the best way to extend the computational aspects of type theory to the specific features of HOTT/UF is an active research area.

symmetry of equality using indiscernibility of identicals become in HoTT/UF constructions of the first level of ∞-groupoid structure.)

In this way, ∞-groupoids become much simpler in HoTT/UF than they are in set theory. We saw in §2 that a mathematician trying to *define* ∞-groupoids in set theory is led to a rather complicated structure. However, HoTT/UF reveals that *synthetically*, an ∞-groupoid is really quite a simple thing: we might say that we obtain a synthetic theory of ∞-groupoids by (0) starting with type theory, (1) taking seriously the idea that a statement of equality $x = y$ should be a type, (2) writing down the most natural rule governing such a type, and then (3) simply *declining to assert* that all such types are mere truth values.[14]

To be precise, however, this is not quite correct; better would be to say that Martin-Löf's type theory, unlike set theory, is sufficiently general to *permit* its types to be treated as ∞-groupoids, and in HoTT/UF we choose to do so. (This is analogous to how intuitionistic logic, unlike classical logic, is sufficiently general to permit the assumption of topological or computational structure.) Thus, in order to obtain a true synthetic theory of ∞-groupoids, we need to add some rules that are *specific* to them, which in particular will ensure that it is definitely *not* the case that all equality types are truth values.

The principal such rule in use is Voevodsky's *univalence axiom* [19]. This is formulated with reference to a *universe type* $\mathcal{U}$, whose points are other types. (For consistency, $\mathcal{U}$ cannot be a point of itself; thus one generally assumes an infinite hierarchy of such universes.) Univalence says that for types $A : \mathcal{U}$ and $B : \mathcal{U}$, the type $A = B$ consists of *equivalences* between A and B, the latter being a standard definition imported from higher category theory[15] that generalizes bijections between sets. In particular, if A has any nontrivial automorphisms, then $A = A$ is not a mere truth value.

Univalence is the central topic of Awodey's chapter; he concludes that it codifies exactly the principle of structuralism, "isomorphic objects are identical". Indeed, with univalence we no longer need any Fregean abstraction to define "structure"; we can simply consider types themselves (or, more generally, types equipped with extra data) to *be* structures. Fregean abstraction is for forgetting irrelevant facts not preserved by isomorphism, like whether $0 \in 1$; but in HoTT/UF there are no such facts, since isomorphic types are actually already *the same*. Thus, if we wish, we may consider HoTT/UF to be a *synthetic theory of structures*.[16]

More concretely, univalence ensures that any construction or proof can be transported across an isomorphism (or equivalence): anything we prove about a type is automatically also true about any equivalent type. Here again HoTT/UF captures precisely an aspect of mathematical practice that is often glossed over by set theory. Univalence also implies that two "truth values", as defined in §4, are equal as soon as they are logically equivalent; thus they really do carry no more information than a truth value.

A second way that the logic of §4 is particularly appropriate is that HoTT/UF clarifies the distinction between types and truth values, by placing it on the first rung of an infinite ladder. In fact, for any integer $n \geq -2$ there is a class of types called *n-types*,[17] such that the singleton is the only (-2)-type, the truth values are the (-1)-types, and the sets are the 0-types. Informally, an n-type contains no higher equality information above level n: two elements of a 0-type (i.e. a set) can be equal in at most one way, two *equalities* in a 1-type can be equal in at most one way, and so on. Formally, A is an n-type if for all $x : A$ and $y : A$, the type $x = y$ is an $(n - 1)$-type (with the induction bottoming out at $n = -2$).

In addition, for any n we have an *n-truncation* operation: $\|A\|_n$ is an n-type obtained from A by discarding all distinctions between equalities above level n. In particular, $\|A\|_{-1}$ discards all distinctions between *points* of A, remembering only whether A is inhabited; thus it is the truth-value truncation $\|A\|$ from §4. The next most important case is the 0-truncation $\|A\|_0$, which makes A into a set by discarding distinctions between equalities between its points, remembering only the truth value of whether or not they are equal.

At this point we can deal with one of the examples of a groupoid from §2: sets and cardinalities. In HoTT/UF the *type of sets* is naturally defined as a subtype of the universe $\mathcal{U}$ which contains only the sets

---

[14]Many people contributed to this view of Martin-Löf's equality types, but Hofmann–Streicher [17] and Awodey–Warren [4] were significant milestones.

[15]Although it requires some cleverness to formulate it correctly in type theory; this was first done by Voevodsky.

[16]This is not in conflict with also calling it a synthetic theory of ∞-groupoids; the two phrases simply emphasize different aspects of HoTT/UF. We could emphasize both aspects at once by calling it a "synthetic theory of ∞-groupoidal structures".

[17]This notion is well-known in homotopy theory under the name *homotopy n-type* and in higher category theory under the name *n-groupoid*. Its definition in type theory is due to Voevodsky, who calls them "types of h-level $n + 2$".

(0-types). By univalence, then, for sets A and B, the type $A = B$ is the type of bijections between them. Thus two sets are *automatically* identical exactly when they are bijective, so it may appear that there is no need to specify the equalities separately from the points in this case.

However, since the type of bijections between sets A and B is itself a set and not (generally) a truth value, the type of sets is a 1-type and not a set. This is an important difference with ZFC, in which the collection of sets (belonging to some universe) *is* itself a set — but it matters little in mathematical practice, which is mostly structural. Indeed, mathematicians familiar with category theory tend to be drawn to this idea: it seems perverse to distinguish between isomorphic sets as ZFC does.[18]

On the other hand, mathematicians *are* accustomed to consider the collection of *cardinalities* to form a set (modulo size considerations). Thus, in HOTT/UF it is sensible to define the set of cardinalities to be the 0-truncation of the type of sets. That is, a cardinality is presented by a set, and bijective sets present equal cardinalities; but unlike sets, two cardinalities can be equal in at most one way. One nice consequence is that the subset of *finite* cardinalities is then equal to the natural numbers.

The 0-truncation has many other uses; for instance, it allows us to import the definition of *homotopy groups* from algebraic topology. Given a type X and a point $x : X$, we first define the *loop space* $\Omega(X, x)$ to be the type $x = x$, and the *n-fold loop space* by induction as $\Omega^{n+1}(X, x) = \Omega^n(\Omega(X, x), \mathsf{refl}_x)$. The nth *homotopy group* of X based at x is then $\pi_n(X, x) = \|\Omega^n(X, x)\|_0$. If X is an n-type, then $\pi_k(X, x)$ is trivial whenever $k > n$; in general it can be said to measure the nontriviality of the identification structure of X at level n. For instance, if X is a set, then $\pi_k(X, x)$ is trivial for any $k \geq 1$; whereas if $X = \mathcal{U}$ and x is a set A, then $\pi_1(\mathcal{U}, A)$ is the automorphism group of A while $\pi_k(\mathcal{U}, A)$ is trivial for $k > 1$.

6 Higher inductive types

I mentioned in §4 that HOTT/UF consists of rules describing operations we can perform on types and their points. In fact, all but a couple of these rules belong to one uniformly specified class, known as *higher inductive types* (HITs), which can be considered a generalization of Bishop's rule for set-construction that takes higher identifications into account.

Higher inductive types include, in particular, *ordinary* inductive types, which have been well-known in type theory for a long time (several examples appear already in [28]). The simplest sorts of these are *nonrecursive*, in which case the rule says that to define a type X, we specify zero or more ways to construct elements of X. This amounts to stipulating some finite list of functions with codomain X and some specified domain, called the *constructors* of X. For instance, given types A and B, their *disjoint union* $A + B$ is specified by saying that there are two ways to construct elements of $A + B$, namely by injecting an element of A or an element of B; thus we have two constructors $\mathsf{inl} : A \to A + B$ and $\mathsf{inr} : B \to A + B$.

As recognized in [28, §1.1], this is similar to Bishop's rule; the main difference is that we omit the specifying of equalities. How then are we to know when two points of such a type are equal? The answer is that an inductive type should be regarded as *freely generated* by its constructors, in the sense that we do not "put in" anything — whether a point or an identification — that is not *forced* to be there by the constructors. For instance, every point of $A + B$ is either of the form $\mathsf{inl}(a)$ or $\mathsf{inr}(b)$, since the constructors do not force any other points to exist. Moreover, no point of the form $\mathsf{inl}(a)$ is equal to one of the form $\mathsf{inr}(b)$, since the constructors do not force any such identifications to exist. However, if we have $a : A$ and $a' : A$ with $a = a'$, then there *is* an induced identification $\mathsf{inl}(a) = \mathsf{inl}(a')$, since all functions (including inl) must respect equality.

More generally, ordinary inductive types can be *recursive*, meaning that some of the constructors of X can take as input one or more previously constructed elements of X. For example, the natural numbers $\mathbb{N}$ have one nonrecursive constructor $0 : \mathbb{N}$ and one recursive one $s : \mathbb{N} \to \mathbb{N}$. The elements and equalities in such a type are all those that can be obtained by applying the constructors, over and over again if necessary.

Higher inductive types are a generalization of ordinary ones, which were invented by the author and

---

[18]This should not be confused with distinguishing between *subsets* of some fixed set that may be abstractly isomorphic as sets, such as $\mathbb{N} \subseteq \mathbb{R}$ and $\mathbb{Q} \subseteq \mathbb{R}$, which is common and essential to mathematics. The point is rather that of Benacerraf [8]: there is no reason to distinguish between, say, $\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}, \dots\}$ and $\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}, \dots\}$ as definitions of "the natural numbers".

others.[19] The simplest case is a nonrecursive *level-1* HIT, where in addition to specifying ways to construct elements of X, we can specify ways to construct identifications between such elements. Thus, in addition to constructor functions as before (which we now call *point-constructors*), we also have *identification-constructors*.

This is almost the same as Bishop's rule for set-construction, with two differences. Firstly, a HIT need not be a set. Secondly, the identification-constructors need not form an equivalence relation; e.g. we may specify $x = y$ and $y = z$ but not $x = z$. However, since all types *are* ∞-groupoids, in such a case it will nevertheless be *true* that $x = z$. More precisely, if we have constructors yielding identifications $p : x = y$ and $q : y = z$, then there will be an induced identification $p \cdot q : x = z$, which is forced to exist even though we didn't "put it in by hand".

Suppose now that we *are* in Bishop's situation, i.e. we have a type A and an equivalence relation $\sim$ on it. We can define a HIT X, with one point-constructor $q : A \to X$, and one identification-constructor saying that whenever $a \sim a'$ we have $q(a) = q(a')$. Then X will be close to the quotient of $\sim$, except that it will not generally be a set even if A is. For instance, since $a \sim a$ for any $a : A$, our identification-constructor yields an identification $q(a) = q(a)$; but nothing we have put into X forces this identification to be the same as $\mathsf{refl}_{q(a)}$, and so (by the free generation principle) it is not. Thus, to obtain the usual quotient of $\sim$, we have to 0-truncate X; in HOTT/UF we may call this the *set-quotient*. For instance, the set of real numbers could be defined as the set-quotient of the equivalence relation on infinite decimal expansions from §2. In this way we essentially recover Bishop's set-formation rule.[20]

Higher inductive types can also be recursive: both kinds of constructor can take previously constructed elements of X as inputs. This is very useful — e.g. it yields free algebraic structures, homotopical localizations, and even the n-truncation — but also somewhat technical, so I will say no more about it.

The reader may naturally wonder *why* we don't ask the identification-constructors to form an equivalence relation. One reason is that for HITs that are not sets, the analogue of an equivalence relation would be an "∞-groupoid" in the exceedingly complicated sense referenced at the beginning of §3. Forcing ourselves to use such structures would vitiate the already-noted advantages of a *synthetic* theory of ∞-groupoids.

As a concrete example of the usefulness of not requiring equivalence relations *a priori*, if we have two functions $f, g : A \rightrightarrows B$ between sets, we can construct their *set-coequalizer* as the 0-truncation of the HIT with one point-constructor $q : B \to X$ and one identification-constructor saying that for any $a : A$ we have $q(f(a)) = q(g(a))$. In set theory, we would have to first construct the equivalence relation on B freely generated by the relations $f(a) \sim g(a)$ and then take its quotient; HITs automate that process for us. Moreover, if we omit the assumption that A and B are sets and also omit the 0-truncation, we obtain a *homotopy coequalizer*, which would be *much* harder to construct otherwise.

Another reason for considering freely generated ∞-groupoids is that many very interesting ∞-groupoids *are* freely generated, and in most cases a fully explicit description of them *is not known* and is not expected to be knowable. Thus, HITs are the *only* way we can represent them in HOTT/UF.

A simple example of a freely generated ∞-groupoid is the *circle*[21] $\mathbb{S}^1$, which as a HIT has one point-constructor $\mathsf{b} : \mathbb{S}^1$ and one identification-constructor $\ell : \mathsf{b} = \mathsf{b}$. Since nothing forces ℓ to be equal to refl_b, it is not — nor is $\ell \cdot \ell$, or $\ell \cdot \ell \cdot \ell$, and so on. In fact, $\Omega(\mathbb{S}^1, \mathsf{b})$ is isomorphic to the integers $\mathbb{Z}$.[22] Since $\mathbb{Z}$ is a

---

[19]Specifically, Lumsdaine, Bauer, and Warren, with further notable contributions by Bruneríe and Licata. The basic theory of HITs is still under development by many people; currently the best general reference is [36, Ch. 6].

[20]There is one subtle difference: Bishop actually allows us to distinguish between 0.5 and $0.4\overline{9}$ as long as we speak of an "operation" rather than a "function". In HOTT/UF such an "operation" is just a function defined on decimal expansions, not anything acting on "real numbers".

[21]This is a "homotopical" circle, not a "topological" circle such as $\{(x, y) \in \mathbb{R} \times \mathbb{R} \mid x^2 + y^2 = 1\}$. The latter can also be defined in HOTT/UF, of course, but it will be a set, whereas the HIT $\mathbb{S}^1$ is not. The homotopical circle is so-called because it is the *shape* (a.k.a. "fundamental ∞-groupoid") of the topological circle, with continuous paths in the latter becoming identifications in the former; and historically ∞-groupoids were originally studied as shapes of topological spaces. In HOTT/UF the shape ought to be constructible as a HIT, but no one has yet managed to do it coherently at all levels. Unlike classically, not every type in HOTT/UF can be the shape of some space, but we can hope that the HIT $\mathbb{S}^1$ is still the shape of the topological circle.

There is an arguably better approach to such questions called "axiomatic cohesion" [34, 35], in which the types of HOTT/UF are enhanced to carry intrinsic topological structure in addition to their higher identifications. Unfortunately, space does not permit me to discuss this here, but a brief introduction can be found in Corfield's chapter.

[22]This is well-known in homotopy theory; its first proof in HOTT/UF by the author [25] was an early milestone in combining HITs with univalence.

set, this implies $\pi_1(\mathbb{S}^1, \mathsf{b}) = \mathbb{Z}$ while $\pi_k(\mathbb{S}^1, \mathsf{b}) = 0$ for all $k > 1$, so in this case we do have a fully explicit description. However, there are similar types for which no such characterization is known, particularly when we move on to *level-n* HITs having constructors of "higher identifications". For instance, the *2-sphere* $\mathbb{S}^2$ has one point-constructor $\mathsf{b} : \mathbb{S}^2$ and one level-2 identification-constructor $\mathsf{refl}_\mathsf{b} = \mathsf{refl}_\mathsf{b}$; the *3-sphere* has $\mathsf{b} : S^3$ with a level-3 $\mathsf{refl}_{\mathsf{refl}_\mathsf{b}} = \mathsf{refl}_{\mathsf{refl}_\mathsf{b}}$; and so on. Analogously to $\mathbb{S}^1$ we have $\pi_n(\mathbb{S}^n) = \mathbb{Z}$,[23] but also for example $\pi_3(\mathbb{S}^2) = \mathbb{Z}$, despite the fact that $\mathbb{S}^2$ has no *constructors* of level 3. In general, $\pi_k(\mathbb{S}^n)$ is usually nontrivial when $k \geq n$, but most of its values are not known. Computing them, for classically defined ∞-groupoids, is a major research area which is not expected to ever be "complete".

What does this mean to a philosopher? For one thing, it shows how a simple foundational system can give rise very quickly to deep mathematics. The rules governing HITs are arguably unavoidable, once we have the idea of defining types in such a way; while the spheres $\mathbb{S}^n$ result from quite simple applications of those rules. Moreover, we have seen that even the basic notion of ∞-groupoid arises inescapably from thinking about equality. Thus, there are numerical invariants like $\pi_3(\mathbb{S}^2)$ quite close to the foundations of logic.

7 General Covariance

At long last, we return to the third example from §2: spacetime manifolds. For simplicity, I will consider only *Minkowski* spacetimes, corresponding to special rather than general relativity; similar ideas can be applied to other kinds of gauge invariance/covariance as well.

A modern mathematician defines a Minkowski spacetime to be a 4-dimensional real affine space with a Lorentzian inner product. We can repeat this definition in HOTT/UF, yielding a type Mink whose points are Minkowski spacetimes. Now we can ask what the *identifications* are in Mink. This is a special case of a more general question: what are the identifications in a *type of structured sets*? Recall that univalence ensures that identifications in the type of *all* sets are bijections; this turns out to imply that an identification of structured sets is a bijection which "preserves all the structure", i.e. an *isomorphism* in the appropriate category (see e.g. [36, §9.8]). Thus, an identification in Mink is an isometry, as we would hope. In particular, anything we can say in HOTT/UF about Minkowski spacetimes is automatically covariant under isometry.

Note that since isometries form a set, Mink is a 1-type. We could, if we wished, 0-truncate it to obtain a set, as we did with the type of sets in §5 to obtain the set of cardinalities. However, the hole argument tells us that this would be *wrong*, at least for the purpose of modeling reality: we really do need to remember the nontrivial identifications in Mink.

So far, so good. However, there is another side to the story, which I alluded to briefly in §2: why did the hole argument seem paradoxical for so long? This can be attributed at least partly to a radically different viewpoint on manifolds, as described by Norton [30]:

> . . . our modern difficulty in reading Einstein literally actually stems from a change. . . in the mathematical tools used. . . . In recent work. . . we begin with a very refined mathematical entity, an abstract differentiable manifold. . . . We then judiciously add further geometric objects only as the physical content of the theory warrants. . . . In the 1910s, mathematical practices in physics were different. . . . one used number manifolds — $\mathbb{R}^n$ or $\mathbb{C}^n$ for example. Thus Minkowski's 'world'. . . was literally $\mathbb{R}^4$, that is it was the set of all quadruples of real numbers.
>
> Now anyone seeking to build a spacetime theory with these mathematical tools of the 1910s faces very different problems from the ones we see now. Modern differentiable manifolds have too little structure and we must add to them. Number manifolds have far too much structure. . . the origin $\langle 0, 0, 0, 0 \rangle$ is quite different from any other point, for example. . . . The problem was not how to add structure to the manifolds, but how to deny physical significance to existing parts of the number manifolds. How do we rule out the idea that $\langle 0, 0, 0, 0 \rangle$ represents the preferred center of the universe. . . ?

In brief, *mathematical structuralism* had not yet been invented. Our explanation of the hole argument relied on comfort with the structural idea of an isometry between abstract manifolds. But if one views spacetime

---

[23] Also a standard result in homotopy theory; see [36, Ch. 8] and [24] for proofs in HOTT/UF.

as the *specific* manifold $\mathbb{R}^4$, this sort of argument is unavailable; thus the confusion surrounding the hole argument becomes more understandable.

While structuralism is the modern method of choice to deal with this conundrum, it is not the only possible solution; historically, Klein's *Erlangen* program was used for the same purpose. Here is Norton again:

> Felix Klein's *Erlangen* program provided precisely the tool that was needed. One assigns a characteristic group to the theory.... Only those aspects of the number manifold that remain invariant under this group are allowed physical significance.... As one increases the size of the group, one strips more and more physical significance out of the number manifold.

This suggests a different definition of Mink: we could begin with the singleton type $\{\mathbb{R}^4\}$ and *add identification-constructors* making up the desired symmetry group (in this case, the Poincaré group[24]). In other words, we say that there is *one* Minkowski spacetime, namely $\mathbb{R}^4$, and that it can be identified with itself in many ways, such as translations, 3D rotations, and Lorentz boosts. These extra added identifications force everything we say about "Minkowski spacetimes" to be invariant under their action. For example, while in $\mathbb{R}^4$ we can distinguish the point $\langle 0, 0, 0, 0 \rangle$, in a Minkowski spacetime we cannot, because this point is not invariant under translations. However, we can say that a Minkowski spacetime comes with a Lorentzian distance function, since this structure on $\mathbb{R}^4$ *is* preserved by the Poincaré group. This is precisely the point of the *Erlangen* program, which HOTT/UF codifies into the foundations of mathematics by constructing a type that "remembers exactly those aspects of $\mathbb{R}^4$ preserved by the group action."

Finally, we can show in HOTT/UF that these two definitions of Minkowski spacetime agree. Roughly, this is because two abstract Minkowski spacetimes can always be identified *somehow*, while their automorphisms can be identified with the Poincaré group; thus the points and the identifications can be matched up consistently. Thus, HOTT/UF could be said to unify the *Erlangen* and structuralist approaches to geometry.

One might argue that these approaches were unified long ago, by the development of category theory. Indeed, as detailed in [27], category theory can be seen as a generalization of the *Erlangen* program, where rather than simply having a group act by automorphisms of a single object, we consider isomorphisms, or more generally morphisms, between different objects, and permit as meaningful only those properties that vary appropriately under such transformations (i.e. those that are covariant — or, perhaps, *contravariant*, the dual sort of variation that can be distinguished only once we allow noninvertible morphisms). And category theory is, of course, the language of choice for the modern structuralist.

However, when category theory is built on top of a foundational set theory, one has to take the additional step of *defining* the notion of isomorphism as the appropriate "criterion of sameness" and (in principle) *proving* that all properties of interest are invariant under isomorphism. As Marquis says, in the *Erlangen* program:

> ... what is usually taken as a *logical* notion, namely equality of objects, is captured in geometry by motions, or transformations of the given group. [27, p19; emphasis added]

Moreover, when generalized to higher groupoids and higher categories, this leads to the highly complicated *defined* notion of ∞-groupoid mentioned in §3. But with univalence and HITs, HOTT/UF places the notion of equality back where it belongs — in logic, or more generally the foundations of mathematics — while maintaining the insights of the *Erlangen* program and category theory.

8   Conclusion

There is much more to HOTT/UF than I have been able to mention in this short chapter, but those aspects I have touched on revolve around a single idea, which generalizes Bishop's set-definition principle: whenever we define a collection of objects, we must also ensure that the identifications and higher identifications between them are correctly specified. Sometimes the correct identifications arise "automatically", such as from the

---

[24]The Poincaré group is usually considered not as a discrete group but as a *Lie* group, with its own manifold structure. This can be incorporated as well using "axiomatic cohesion", mentioned briefly in footnote 21 on page 11.

univalence axiom; other times we have to generate new ones, as with higher inductive types. But in no case must we (or even *can* we) separate those identifications from the objects themselves: with ∞-groupoids as basic foundational objects, every collection carries along with itself the appropriate notion of identification between its objects, higher identification between those, and so on. This can be regarded as the central innovation of HoTT/UF, both for mathematics and for philosophy.

References

[1] Peter Aczel and Michael Rathjen. Notes on constructive set theory. Mathematical Logic 40, Reports Institut Mittag-Leffler, 2000/2001. Available at `http://www.ml.kva.se/preprints/meta/AczelMon_Sep_24_09_16_56.rdf.html`.

[2] Steve Awodey. Type theory and homotopy. In *Epistemology versus ontology*, volume 27 of *Log. Epistemol. Unity Sci.*, pages 183–201. Springer, Dordrecht, 2012.

[3] Steve Awodey, Álvaro Pelayo, and Michael A. Warren. Voevodsky's univalence axiom in homotopy type theory. *Notices Amer. Math. Soc.*, 60(9):1164–1167, 2013.

[4] Steve Awodey and Michael A. Warren. Homotopy theoretic models of identity types. *Math. Proc. Camb. Phil. Soc.*, 146(45):45–55, 2009.

[5] Steven Awodey and Andrej Bauer. Propositions as [types]. *J. Logic Comput.*, 14(4):447–471, 2004.

[6] John Baez. The homotopy hypothesis. `http://math.ucr.edu/home/baez/homotopy/`, January 2007. Lecture at *Higher Categories and Their Applications*.

[7] M. Beeson. *Foundations of Constructive Mathematics*. Springer, 1985.

[8] Paul Benacerraf. What numbers could not be. *The Philosophical Review*, 74(1):47–73, Jan 1965.

[9] E. Bishop. *Foundations of constructive analysis*. McGraw-Hill series in higher mathematics. McGraw-Hill, 1967.

[10] R. L. Constable, S. F. Allen, H. M. Bromley, W. R. Cleaveland, J. F. Cremer, R. W. Harper, D. J. Howe, T. B. Knoblock, N. P. Mendler, P. Panangaden, J. T. Sasaki, and S. F. Smith. *Implementing mathematics with the Nuprl proof development system*. Prentice-Hall, 1986.

[11] H. B. Curry. Functionality in combinatory logic. *Proceedings of the national academy of science*, 20:584–590, 1934.

[12] W. Ewald. *From Kant to Hilbert: a source book in the foundations of mathematics*. Clarendon Press, 1996. In two volumes.

[13] Gottlob Frege. *Die Grundlagen der Arithmetik: eine logisch-mathematische Untersuchung über den Begriff der Zahl*. Breslau: W. Koebner, 1884.

[14] D. Hilbert and P. Bernays. *Grundlagen der Mathematik*. Springer, 1934–1939. (in two volumes).

[15] David Hilbert. *Grundlagen der Geometrie*. Teubner, Berlin, 1899.

[16] David Hilbert. Über den Zahlbegriff. *Jahresbericht der deutschen Mathematiker–Vereinigung*, 8:180–184, 1900. English translation in [12].

[17] Martin Hofmann and Thomas Streicher. The groupoid interpretation of type theory. In *Twenty-five years of constructive type theory (Venice, 1995)*, volume 36 of *Oxford Logic Guides*, pages 83–111. Oxford Univ. Press, New York, 1998.

[18] W. A. Howard. The formulae-as-types notion of construction. In *To H. B. Curry: Essays on Combinatory Logic, Lambda Calculus, and Formalism*, pages 479–491. Academic Press, 1980. Notes originally circulated privately in 1969.

[19] Chris Kapulkin, Peter LeFanu Lumsdaine, and Vladimir Voevodsky. The simplicial model of univalent foundations. arXiv:1211.2851, 2012.

[20] F. William Lawvere. The category of categories as a foundation for mathematics. In *Proc. Conf. Categorical Algebra (La Jolla, Calif., 1965)*, pages 1–20. Springer, New York, 1966.

[21] F. William Lawvere. Equality in hyperdoctrines and comprehension schema as an adjoint functor. In *Applications of Categorical Algebra (Proc. Sympos. Pure Math., Vol. XVII, New York, 1968)*, pages 1–14. Amer. Math. Soc., Providence, R.I., 1970.

[22] F. William Lawvere. An elementary theory of the category of sets (long version) with commentary. *Repr. Theory Appl. Categ.*, 11:1–35 (electronic), 2005. Reprinted and expanded from Proc. Nat. Acad. Sci. U.S.A. **52** (1964), With comments by the author and Colin McLarty.

[23] William Lawvere. Cohesive toposes and Cantor's "lauter Einsen". *Philosophia Mathematica*, 2(3):5–15, 1994.

[24] Daniel R. Licata and Guillaume Brunerie. $\pi_n(S^n)$ in homotopy type theory. *CPP*, 2013. `http://dlicata.web.wesleyan.edu/pubs/lb13cpp/lb13cpp.pdf`.

[25] Daniel R. Licata and Michael Shulman. Calculating the fundamental group of the circle in homotopy type theory. In *LICS'13*, 2013.

[26] Peter LeFanu Lumsdaine. Weak omega-categories from intensional type theory. *Typed lambda calculi and applications*, 6:1–19, 2010.

[27] Jean-Pierre Marquis. *From a geometrical point of view: a study of the history and philosophy of category theory*. Springer, 2009.

[28] Per Martin-Löf. An intuitionistic theory of types: predicative part. In *Logic Colloquium*. North Holland, 1975.

[29] Martín Hözel Escardó and Chuangjie Xu. The inconsistency of a Brouwerian continuity principle with the Curry–Howard interpretation. *Typed Lambda Calculi and Applications*, 2015.

[30] John Norton. General covariance and the foundations of general relativity: eight decades of dispute. *Rep. Prog. Phys.*, 56:791–858, 1993.

[31] Álvaro Pelayo and Michael A. Warren. Homotopy type theory and Voevodsky's univalent foundations. *Bull. Amer. Math. Soc. (N.S.)*, 51(4):597–648, 2014.

[32] Andrei Rodin. On constructive axiomatic method. arXiv:1408.3519, 2014.

[33] R. K. Sachs and H. Wu. *General relativity for mathematicians*. Springer–Verlag, 1977.

[34] Urs Schreiber and Michael Shulman. Quantum gauge field theory in cohesive homotopy type theory. In *QPL'12*, 2012. `http://ncatlab.org/schreiber/files/QFTinCohesiveHoTT.pdf`.

[35] Michael Shulman. Brouwer's fixed-point theorem in real-cohesive homotopy type theory. arXiv:1509.07584, 2015.

[36] Univalent Foundations Program. *Homotopy Type Theory: Univalent Foundations of Mathematics*. `http://homotopytypetheory.org/book/`, first edition, 2013.

[37] Benno van den Berg and Richard Garner. Types are weak ω-groupoids. *Proceedings of the London Mathematical Society*, 102(2):370–394, 2011.

[38] Philip Wadler. Propositions as types. *Communications of the ACM*, 2015. `http://homepages.inf.ed.ac.uk/wadler/papers/propositions-as-types/propositions-as-types.pdf`.