Set-theoretic universes and paradoxes in elementary 2-topoi

Joj Helfer

June 3, 2026

Simons Center for Geometry and Physics

Overview

- Goals:
  - Describe a foundational theory based on the 2-category of categories
  - Construct models of set theory using this framework
- Background:
  - Topos theory
  - 2-categories
  - Models of set theory
- Foreground:
  - 2-topos theory
  - Categorical models of set theory

Topoi

Definition of Topos

Definition

A(n elementary) **topos** is a category $\mathbf{C}$ which has

- finite limits (a terminal object and pullbacks)
- exponential objects Y^X
- a subobject classifier Ω

Equivalently, $\mathbf{C}$ has

- finite limits
- power objects $\mathcal{P}X$

Most important example: $\mathbf{C} = \mathbf{Set}$

Subobject classifier

The subobject classifier Ω **represents the subobject functor**:

$$\mathbf{C}(A, \Omega) \cong \mathsf{Sub}(A)$$

where

$$\mathsf{Sub}(A) := \{P \rightarrowtail A\}/\cong.$$

- In **Set**: $\Omega = \{0, 1\}$
- Ω is also called the **object of truth values**
- There are always **elements** $\mathsf{true}, \mathsf{false} \colon \mathbf{1} \to \Omega$ representing the subobjects $\mathrm{id}_{\mathbf{1}} \colon \mathbf{1} \rightarrowtail \mathbf{1}$ and $\emptyset \rightarrowtail \mathbf{1}$.

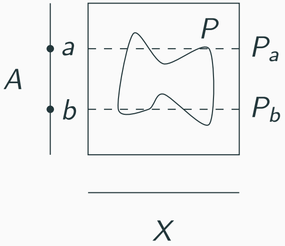
Power object

The power object $\mathcal{P}X$ **represents the families-of-subobjects-of-X functor**:

$$\mathbf{C}(A, \mathcal{P}X) \cong \mathrm{Sub}_X(A)$$

where

$$\mathrm{Sub}_X(A) := \{P \rightarrowtail A \times X\}/\cong.$$



Relationship to Ω:

- $\Omega \cong \mathcal{P}(\mathbf{1})$
- $\mathcal{P}(X) \cong \Omega^X$

Natural numbers objects

We often assume a topos also has a **natural numbers object** N:

This is equipped with a **universal diagram** $\mathbf{1} \xrightarrow{0} N \xrightarrow{s} N$

$$
\begin{tikzcd}
 & N \arrow[r, "s"] \arrow[dd, dashed, "F", "\exists!"'] & N \arrow[dd, dashed, "F"] \\
\mathbf{1} \arrow[ur, "0"] \arrow[dr, "a", "\forall"'] & & \\
 & A \arrow[r, "f", "\forall"'] & A
\end{tikzcd}
$$

In **Set**: $N = \mathbb{N}$ and $F(n) = f^n(a)$.

Examples of topoi

- **Set**
- $\mathbf{Set}_{\mathrm{fin}}$ (no natural numbers object)
- The category $\mathbf{Set}^{\mathbf{C}^{\mathrm{op}}}$ of **presheaves** on any small category $\mathbf{C}$
- More generally, the category $\mathrm{Sh}(\mathbf{C})$ of **sheaves** on any small **site** $\mathbf{C}$

Interpreting logical formulas in a topos $\mathbf{C}$: subobjects

Example:

Fix

- an object $X \in \mathbf{C}$
- a morphism $f\colon X \to X$
- a binary relation $\prec\ \rightarrowtail X \times X$

Let $\varphi(x)$ be the formula

$$x \prec f(x).$$

We can then define a subobject

$$A = \text{“}\{x \in X \mid \varphi(x)\}\text{”} \rightarrowtail X.$$

For a **generalized element** $x\colon U \to X$, we then have “$x \in A$” iff “$x \prec f(x)$”:

$$\begin{tikzcd} & A \arrow[d, tail] \\ U \arrow[ur, dashed] \arrow[r, "x"'] & X \end{tikzcd} \quad \text{iff} \quad \begin{tikzcd} & \prec \arrow[d, tail] \\ U \arrow[ur, dashed] \arrow[r, "\langle x, f\circ x\rangle"'] & X \times X. \end{tikzcd}$$

We can do this with “**any**” logical formula $\varphi(x)$ in place of $\varphi(x) \Leftrightarrow x \prec f(x)$.

Interpreting logical formulas in a topos $\mathbf{C}$: truth values

Example:

Fix

- an object $X \in \mathbf{C}$
- a morphism $f\colon X \to X$
- a binary relation $\prec \rightarrowtail X \times X$

Consider the **closed** formula

$$\varphi \equiv \forall x, y \in X.\ x \prec y \Rightarrow f(x) \prec f(y).$$

- This (or any closed) formula has a **truth value** $[\varphi]\colon \mathbf{1} \to \Omega$ (equivalently, $[\varphi] \rightarrowtail \mathbf{1}$)
- We have $[\varphi] = \mathrm{true}\colon \mathbf{1} \to \Omega$ iff for all $x, y\colon U \to X$, "$x \prec y$" implies "$f(x) \prec f(y)$":

$$\begin{tikzcd} & \prec \arrow[d, tail] \\ U \arrow[ur, dashed] \arrow[r, "\langle x,y \rangle"'] & X \times X \end{tikzcd} \quad \Longrightarrow \quad \begin{tikzcd} & \prec \arrow[d, tail] \\ U \arrow[ur, dashed] \arrow[r, "\langle f\circ x, f\circ y \rangle"'] & X \times X. \end{tikzcd}$$

Interpreting logical formulas in a topos $\mathbf{C}$

Other closed formulas φ one might consider:

(i) The fundamental theorem of arithmetic

(ii) The continuum hypothesis

(iii) The axiom of choice

(iv) "Every function $f : [0, 1] \to \mathbb{R}$ is continuous" (Brouwer's theorem)

(v) "Every function $f : \mathbb{N} \to \mathbb{N}$ is computable" (Church-Turing thesis)

In the case (i), φ holds in **every** topos $\mathbf{C}$

- For (ii) and (iii), there exist topoi $\mathbf{C}$ where φ holds, and others where φ fails.
- The same goes for (iv) and (v)!

Topos theory as a foundational theory

In an axiomatic system, we have **undefined/primitive notions** and **axioms**.

- In **set theory**, the primitive notions are **set** and the **membership relation** $x \in y$.
- In **topos theory**, the primitive notions are **set**, **function** $f : x \to y$ and **composition** $g \circ f$.
  - And the **axioms** are very simple: associativity, identities, finite limits and power objects (and NNO)!

The relationship between these foundational theories is well-studied, and we will return to it later.

2-topoi

The basic idea

"2-topos" : $\mathcal{C}at$ :: "topos" : **Set**

Enriched categories

$\mathcal{V}$: a category with a finite products.

A **$\mathcal{V}$-enriched category** $\mathcal{C}$ has:

- a set $\mathrm{Ob}\,\mathcal{C}$ of objects
- an hom-object $\mathcal{C}(X, Y) \in \mathcal{V}$ for each $X, Y \in \mathcal{C}$
- a composition morphism $\mathcal{C}(X, Y) \times \mathcal{C}(Y, Z) \to \mathcal{C}(X, Z)$ in $\mathcal{V}$ for all $X, Y, Z \in \mathcal{C}$.

(Examples: additive categories*, topological categories.)

- **"2-category"** $=$ category enriched in **Cat**, the 1-category of categories.
- **Main example**: $\mathcal{C}at$, the 2-category of categories
- $\mathcal{C}at(X, Y) \in$ **Cat**: the usual functor category

2-categories

More directly, a 2-category $\mathcal{C}$ has:

- **objects** X, Y, Z
- **morphisms** $X \xrightarrow{f} Y \xrightarrow{g} Z$
- **2-cells** $$\begin{tikzcd} X \arrow[r, bend left, "f"] \arrow[r, bend right, "g"'] \arrow[r, phantom, "\Downarrow \alpha"] & Y \arrow[r, bend left, "h"] \arrow[r, bend right, "k"'] \arrow[r, phantom, "\Downarrow \beta"] & Z \end{tikzcd}$$
- **composition** of morphisms $X \xrightarrow{g \circ f} Z$
- **horizontal composition** $$\begin{tikzcd} X \arrow[r, bend left, "h \circ f"] \arrow[r, bend right, "k \circ g"'] \arrow[r, phantom, "\Downarrow \beta\circ\alpha"] & Z \end{tikzcd}$$
- **vertical composition** $$\begin{tikzcd} X \arrow[r, bend left, "f"] \arrow[r, "g"] \arrow[r, bend right, "\ell"'] & Y \end{tikzcd} \quad \mapsto \quad \begin{tikzcd} X \arrow[r, bend left, "f"] \arrow[r, bend right, "\ell"'] \arrow[r, phantom, "\Downarrow \gamma\circ\alpha"] & Y \end{tikzcd}$$
- **laws**

Categorical foundations

Another idea for an axiomatic system (Lawvere '65, Weber '07, Makkai '90s):

The primitive notions are: **category**, **functor**, **natural transformation**, and the three kinds of **composition**

- What are the right **axioms**?
- In this theory, we would want to be able to talk about the **category of sets**.

The DOF classifier

Idea (Lawvere, Weber):

$$\mathbf{Set} : \mathcal{C}at :: \Omega : \mathbf{Set}.$$

- We saw that Ω classifies **subobjects**: $\mathbf{C}(X, \Omega) \cong \mathrm{Sub}(X)$.
- What does $\mathbf{Set} \in \mathcal{C}at$ classify?
  - Answer: **Discrete OpFibrations**:

$$\mathcal{C}at(X, \mathbf{Set}) \cong \mathrm{DOF}(X).^*$$

DOFs

A functor $p\colon E \to B$ is a **DOF** if:

$$\begin{tikzcd}
E \arrow[d, "p"] & \ni & \forall \tilde{x} \arrow[r, dashed, "\exists ! \tilde{f}"] & \exists ! \tilde{y} \\
B & \ni & \forall x \arrow[r, "\forall f"] & \forall y
\end{tikzcd}$$

- Given $f\colon x \to y$, we get $f_*\colon p^{-1}(x) \to p^{-1}(y)$ with $f_*(\tilde{x}) = \tilde{y}$.
- This gives a functor $B \to \mathbf{Set}$.

2-topoi

- There is no settled definition of elementary 2-topos.
- Weber gave some axioms, and I give some more (following Boshuk-Makkai).
- Most important is **existence of a DOF classifier** $\mathbb{S} \in \mathcal{C}$.
- There are still lots of **examples**; e.g., $\mathcal{C}at^{\mathbf{C}}$ for any $\mathbf{C}$.
- (Sweeping some things under the rug...)

Theorem (H. 2024, paraphrased)

The DOF classifier $\mathbb{S}$ *in a 2-topos* $\mathcal{C}$ *is an* **internal 1-topos***.*
In particular, the 1-category $\mathcal{C}(\mathbf{1}, \mathbb{S})$ *is a 1-topos.*

- Slogan: "2-topos theory subsumes topos theory"
- Goal: show that 2-topos theory subsumes **set theory**

Set theory

Abstract and concrete sets

- The sets (objects) X in topos theory are **abstract sets** (Lawvere '76, Cantor 1895, Makkai '13)
    - There are no **things** x in the theory for which we can ask "is x in X?"
- By contrast, the sets in set theory satisfy **extensionality**:
$$X = Y \iff \forall x.\,(x \in X \leftrightarrow x \in Y).$$
- Standard axiomatic set theories (e.g., ZF) deal specifically with **pure** or **hereditary** sets
- A set is **pure** if all of its elements are pure sets (recursive definition).

Examples of pure sets

(0) $a := \{\}$

(1) $b := \{a\} = \{\{\}\}$

(2) $c := \{b\} = \{\{\{\}\}\}$ and
$d := \{a, b\} = \{\{\}, \{\{\}\}\}$

(3) $\{c\}$, $\{d\}$, $\{c, d\}$, $\{a, c\}$, $\{a, d\}$,
$\{a, c, d\}$, $\{b, c\}$, $\{b, d\}$, $\{b, c, d\}$,
$\{a, b, c\}$, $\{a, b, d\}$, $\{a, b, c, d\}$

(4) etc.

- At each stage, we form sets using sets from the previous stages.
- These stages are called the **cumulative hierarchy** or **von Neumann hierarchy** V_α.
- $\mathrm{V} = \bigcup_\alpha \mathrm{V}_\alpha$ is the **universe of pure sets**.

Pure sets as abstract structures

- A pure set x is **transitive** if
$$z \in y \in x \to z \in x \text{ for all } z, y$$
- The **transitive closure** $\mathrm{tc}(x)$ of a pure set x is the least transitive superset of x
- Example:
$$\mathrm{tc}\left(\left\{\{\{\emptyset\}\}\right\}\right) = \left\{\{\{\emptyset\}\}, \{\emptyset\}, \emptyset\right\}$$
- Given a pure set x, equip $\mathrm{tc}(\{x\})$ with the binary relation $\prec_x$ given by
$$z \prec_x y \iff z \in y.$$

Fact (Mostowski '49, Cole '72, Osius '73)

If $(\mathrm{tc}(\{x\}), \prec_x) \cong (\mathrm{tc}(\{y\}), \prec_y)$, *then* $x = y$.

Pure sets as abstract structures, continued

Proposition (Mostowski collapsing lemma, '49)

Let $(X, \prec)$ be a set with a binary relation. Then $(X, \prec) \cong (\mathrm{tc}(\{x\}), \prec_x)$ for some (unique) pure set x if and only if:

(i) $\prec$ *is* **extensional**:
$$(\forall c.\ c \prec a \leftrightarrow c \prec b) \to a = b.$$

(ii) $\prec$ *is* **well-founded**: *any* $\emptyset \neq S \subset X$ *has a* $\prec$*-minimal element.*

(iii) X *has a* **top element** t_X: *if* $S \subset X$ *contains* t_X *and is* **downward closed**:
$$a \prec b \in S \to a \in S,$$
then $S = X$.

Let us call a structure $(X, \prec)$ satisfying (i)-(iii) a **Mostowski structure**.

Pure sets as abstract structures, concluded

Conclusion:

The map

$$\mathrm{tc}\colon \mathrm{V} \to \mathcal{M}$$

from the class V of pure sets to the class $\mathcal{M}$ of (isomorphism classes of) Mostowski structures is an **isomorphism**.

- Slogan: pure-set theory is part of abstract set theory; it is the study of certain **abstract structures** (Mostowski structures)

Mostowski structures in topoi

A **Mostowski structure in a topos C** consists of:

- an object $X \in \mathbf{C}$, and
- a binary relation $\prec \rightarrowtail X \times X$
- satisfying (i)-(iii) as in Mostowski's lemma.

This allows us to **interpret** pure-set theory into topos theory:

- "**set**" (in pure-set theory) is interpreted as "**Mostowski structure**" (in topos theory)
- Under additional assumptions on the topos **C**, this interpretation has been proven (Cole, Osius, '72) to satisfy a **weak version of ZF**.
- Under different additional assumptions on **C** ("category of classes"), Joyal-Moerdijk '95 directly construct a model of **full ZF*** in **C**.
- See also `www.phil.cmu.edu/projects/ast/` & Shulman's *Stack Semantics* '10.

Conclusion

Theorem (H., paraphrased)

Any 2-topos $\mathcal{C}$ contains a model $\mathcal{V}$ of full ZF set theory.*

- The proof combines elements of both the “Mostowski structures in a topos” approach and the “category of classes” approach.
- Also closely related to Shulman’s ideas in “Stack Semantics” and elsewhere.

Bonus slide 1: arrow objects and cotensors

Given an object X in a 2-category $\mathcal{C}$, an **arrow object** $X^{\rightarrow}$ is determined by the universal property

$$\mathcal{C}(A, X^{\rightarrow}) \cong \mathcal{C}(A, X)^{\rightarrow}$$

for any $A \in \mathcal{C}$.

More generally, for any $X \in \mathcal{C}$ and small category J, the **cotensor** X^J of X by $\mathcal{C}$ satisfies

$$\mathcal{C}(A, X^J) \cong \mathcal{C}(A, X)^J.$$

Fact

If a 2-category $\mathcal{C}$ has finite limits and arrow objects, it has all cotensors X^J by finite categories J.

Bonus slide 2: cotensors by higher-order theories

A **second-order theory** is given by a set of sentences concerning the elements of a set X, given functions and relations on X, and subsets of X.

An example is the theory of Mostowski structures.

For any finite second-order theory $\mathcal{T}$ and any topos $\mathbf{C}$, we can form the groupoid $\mathbf{C}^{\mathcal{T}}$ of models of $\mathcal{T}$ in $\mathbf{C}$.

Theorem (H., paraphrased)

The DOF classifier $\mathbb{S}$ *in any 2-topos* $\mathcal{C}$ *admits a cotensor* $\mathbb{S}^{\mathcal{T}}$ *by any finite second-order theory* $\mathcal{C}$*: this satisfies*

$$\mathcal{C}(\mathbf{1}, \mathbb{S}^{\mathcal{T}}) \cong \mathcal{C}(\mathbf{1}, \mathbb{S})^{\mathcal{T}}.$$

We can thus define the internal object $\mathcal{M}(\mathbb{S}) \in \mathcal{C}$ of Mostowski structures in $\mathbb{S}$.

Bonus slide 3: size issues

It is not quite true that

$$\mathcal{C}at(X, \mathbf{Set}) \cong \mathsf{DOF}(X).$$

Rather, functors $X \to \mathbf{Set}$ correspond only to **fibrewise small** DOFs $p \colon E \to X$: those for which each fibre $p^{-1}(x)$ is small.

Thus, in a 2-topos $\mathcal{C}$, we can only demand that the canonical morphism $\mathcal{C}(X, \mathbb{S}) \to \mathsf{DOF}(X)$ is **fully faithful**, rather than an equivalence. Indeed:

Theorem (H., paraphrased)

If a 2-topos $\mathcal{C}$ contains a DOF classifier $\mathbb{S}$ such that $\mathcal{C}(X, \mathbb{S}) \to \mathsf{DOF}(X)$ is an equivalence for all X, then $\mathcal{C}$ is trivial.

Bonus slide 4: discreteness of $\mathcal{M}(\mathrm{S})$ and categories of classes

We can prove that $\mathcal{M}(\mathrm{S})$ is **discrete**: the morphism $\mathcal{M}(\mathrm{S}) \to \mathbf{1}$ is a DOF.

Under the assumption that S classifies **all** DOFs, $\mathcal{M}(\mathrm{S})$ would thus be an object in $\mathsf{DOF}(\mathbf{1}) \cong \mathcal{C}(\mathbf{1}, \mathrm{S})$. We thus have that "the universe of all sets is a set" (rather than a class); this is what leads to a Russell-type contradiction and proves the above theorem.

Instead, we assume that $\mathcal{M}(\mathrm{S}) \to \mathbf{1}$ is **small** with respect to a second DOF classifier S'.

We thus get $\mathcal{M}(\mathrm{S})$ as an object of the topos $\mathbf{C} := \mathsf{DOF}_{\mathrm{S}'}(\mathbf{1}) \cong \mathcal{C}(\mathbf{1}, \mathrm{S}')$ of S'-small discrete objects.

This topos is equipped with a special sub-topos consisting of the S-small discrete objects (it is something like a "category of classes").

The object $\mathcal{M}(\mathrm{S}) \in \mathbf{C}$ has a special property: it classifies **S-small Mostowski structures**. I prove that any such object is a model for ZF in $\mathbf{C}$.