Some context for the recent proof of the consistency of Quine's set theory New Foundations

Lavinia Randall Holmes
Boise State University (emeritus)

March 6, 2026

Abstract

The author, with Sky Wilshaw, has recently completed a proof of the consistency of Quine's set theory New Foundations (NF), which has been formally verified in Lean by Sky Wilshaw. This talk will not present this entire proof (this cannot be done in 45 minutes) but will discuss the context from which this proof comes and perhaps some general information about the approach taken.

We will tell the story from the beginning, and articulate the problem at the appropriate place in the narrative, and explain what we did at the appropriate place in the narrative. Hereinafter, “I/me” denotes Randall Holmes and “we” denotes one or both of Randall Holmes and Sky Wilshaw.

The story begins with type theory. The very beginning of the story we will not tell. The first system of type theory proposed is the (not entirely specified) system in Russell and Whitehead, *Principia Mathematica*. Merely describing this system would raise enough questions for a 45 minute talk.

The system of type theory which is the parent of New Foundations was originally proposed by Tarski in about 1930.

The name for it we use, TST (*théorie simple des types*) was introduced by the school of Belgian logicians who studied NF.

Uninformed accounts of Russell's type theory are often actually descriptions of TST.

TST is a first order theory with sorts indexed by the natural numbers $0, 1, 2, \ldots$.

We do not follow the original practice of indexing all our variables with type superscripts. We say that each variable x in our language has a sort $\texttt{type}(x)$ (a natural number) and that we have a countable infinity of variables of each type. Atomic formulas $x = y$ are well formed iff $\texttt{type}(x) = \texttt{type}(y)$. We do nod to the superscripting practice by stipulating that $\texttt{type}(x^i) = i$: if a variable actually has a superscript, this indicates its type. Atomic formulas $x \in y$ are well-formed iff $\texttt{type}(x) + 1 = \texttt{type}(y)$.

The axioms of TST look exactly like the axioms of naiv̈e set theory, to a naïve eye.

Each formula of the shape
$(\forall xy : x = y \leftrightarrow (\forall z : z \in x \leftrightarrow z \in y))$ is an axiom. These are extensionality axioms. This is a countably infinite scheme (up to renaming of bound variables) with one axiom for each type to be assigned to x.

Each formula of the shape
$(\exists A : (\forall x : x \in A \leftrightarrow \phi))$ in which A is not free in ϕ is an axiom. These are comprehension axioms. The disastrous Russell class instance does not occur because there is no formula $x \not\in x$.

We introduce terms $\{x : \phi\}$ (of the same type as A) to denote the unique witness to the axiom $(\exists A : (\forall x : x \in A \leftrightarrow \phi))$.

The informal picture of what is going on here is that type 0 is inhabited by objects of an entirely unspecified nature which we call “individuals” and type $n + 1$ is inhabited by collections of inhabitants of type n.

The axioms say that the identity criterion on objects of positive type is what we expect (extensionality) and that each property of type n objects that we can express is actually implemented as a type $n+1$ object (comprehension).

A model of TST in ordinary set theory is obtained if we specify a set X_0 to implement the type of individuals, and, once we have implemented type n as a set X_n, implement type $n + 1$ as $X_{n+1} = \mathcal{P}(X_n)$. Membership and equality relations of the model are appropriate subsets of the membership and equality relations of the metatheory. Of course, this implements the informal description above.

TST is fairly competent mathematically if one adds axioms of infinity and choice. Historically, the types have been regarded as cumbersome and untyped theories have been more popular.

We discuss the implementation of the set of natural numbers, both to illustrate ways to do math in this system, and to make another point which leads us into the story of New Foundations.

The idea is to define each natural number n as the set of all sets with n elements. This may appear circular. It is not (or the circularity can be avoided).

We introduce some familiar notation. $\emptyset^{i+1} = \{x^i : x^i \neq x^i\}$. $\{x, y\} = \{z : z = x \vee z = y\}$. $\{x\} = \{x, x\}$. $A \cup B = \{x : x \in A \vee x \in B\}$. For $n > 2$, $\{x_1, x_2, \ldots x_n\} = \{x_1\} \cup \{x_2, \ldots, x_n\}$ (a definition by recursion on n in the metatheory).

We define 0^{i+2} as $\{\emptyset^{i+1}\}$. We note that this is the type $i+2$ set of all type $i+1$ sets with 0 elements.

We denote $\{a \cup \{x\} : a \in A \wedge x \not\in a\}$ by $\sigma(A)$, for any set A. Notice that if N is the set of all type $i+1$ sets with n elements, $\sigma(N)$ is the set of all type $i+1$ sets with $n+1$ elements. Thus $\sigma(0^{i+2})$, which we call 1^{i+2}, is the set of all type $i+1$ sets with 1 element, and $\sigma(1^{i+2})$, which we call 2^{i+2}, is the set of all type $i+1$ sets with two elements, and so forth.,

The form of the set abstract with a complex term to the left of the colon may require comment: $\{t(x_1, \ldots, x_n) : \phi\}$ can be read as $\{y : (\exists x_1, \ldots, x_n : y = t(x_1, \ldots, x_n) \wedge \phi)\}$, where y does not occur in ϕ.

We define $\mathtt{Ind}^{i+4}$, the set of inductive sets, as $\{I : 0^{i+2} \in I \wedge (\forall n : n \in I \rightarrow \sigma(n) \in I)\}$.

Now we can define $\mathbb{N}^{i+3}$ as $\{n : (\forall I : I \in \mathtt{Ind}^{i+4} \rightarrow n \in I\}$. Further we can define $\mathtt{Fin}^{i+2}$, the collection of finite sets, as $\{A : (\exists n \in \mathbb{N}^{i+3} : A \in n)\}$.

Further, we can articulate the axiom (scheme) of Infinity. We define V^{i+1} as $\{x^i : x^i = x^i\}$ and state the axiom (scheme) of infinity as $\neg V^{i+1} \in \mathtt{Fin}^{i+2}$.

TST + Infinity is strong enough to do all mathematics outside technical set theory. One would probably want the axiom (scheme) of choice (every partition has a choice set), which can be stated in basically the same way as in ordinary set theory.

We think that the statement and development of this definition of the natural numbers (ultimately due to Frege) is really interesting, but the exposition of specific mathematics is not really our aim here.

Notice that we did not implement a single set of natural numbers. We implemented a distinct set of natural numbers in each type. Notice also that while we have avoided encumbering our notation with type superscripts on variables, we have had to put type superscripts on constants we have defined. In fact, we could omit these type superscripts (and developments of mathematics in type theory have been given in which the superscripts are left implicit) because in general the types of constants can be deduced from the context in which they appear.

But a statement such as $\neg V \in \texttt{Fin}$, the axiom of infinity in a superscript-free form, strains this convention because there is no variable in sight from which to deduce the types of the constants. We must be asserting this for *every* type and our language does not actually support quantification over types. Such generality can be made rigorous (we can explain that this is what we mean, but it is tricky).

It would be an interesting project (which we have undertaken elsewhere) to make rigorous a superscript free style of doing mathematics in TST. But that is not where we are going today.

This is where the story of New Foundations starts. Quine observed the phenomena we are describing and observed that TST has a kind of symmetry, called “typical ambiguity” in the NF literature (Russell used the term “systematic ambiguity” for an analogous phenomenon in the much more complex type system of *Principia Mathematica*.) We are going to present this a bit anachronistically, using terminology introduced later by Specker, for efficiency.

We postulate a map ($x \mapsto x^+$) on our variables, whose restriction to variables of each type i is a bijection from the type i variables to the type $i+1$ variables. For any formula ϕ, we define ϕ^+ as the formula obtained by replacing each variable x in ϕ with x^+.

Without using this notation [but it is much easier to say this way] Quine observed that for each axiom ϕ of TST, ϕ^+ is also an axiom (evident by inspection) and further that for each theorem ϕ of TST, ϕ^+ is also a theorem. Each object that we define by a set abstract $\{x : \phi\}$ has an exact analogue $\{x^+ : \phi^+\}$ in the next higher type, and in fact analogues in each higher type obtained by iterating this process. The entire world of TST, in terms of what we can prove and what we can define, looks like a hall of mirrors.

Quine then made a bold move. He proposed that the types are surplus to requirements, and we should view all the types as the same, ϕ and ϕ^+ as the same statement (when ϕ is a closed formula of course) and $\{x : \phi\}$ as the same object as $\{x^+ : \phi^+\}$. Under this proposal, we do not have a series of empty sets $\emptyset^{i+1}$ concerning which we cannot even say whether they are the same or different: we have one empty set. We do not have a series of sets V^{i+1} each of which is the set of all type i objects: we have one set V which is the set of *all* objects. We do not have a Frege natural number 3^{i+2} in each type $i+2$ which is the set of all sets of three type i objects: we are back to Frege's own magnificent 3, the set of all sets with three elements.

The theory which formalizes Quine's bold proposal (called NF, or New Foundations, after the title "New foundations for mathematical logic" of the paper in which Quine proposed this in 1937) is as follows.

NF is a first order unsorted theory with an axiom of extensionality
$(\forall xy : x = y \leftrightarrow (\forall z : z \in x \leftrightarrow z \in y))$ [this is a single axiom, not a scheme] and the axioms of comprehension $(\exists A : (\forall x : x \in A \leftrightarrow \phi))$ which are transcriptions of comprehension axioms of TST, in the sense that there is an bijection from variables of TST to variables of NF and a formula ϕ^* of the language of TST such that replacing each variable in ϕ^* with its image under the bijection gives ϕ.

Quine didn't put it this way. It is inelegant to pose the axioms of one theory in terms of the language and axioms of another. He defined a condition "ϕ is stratified" and asserted $(\exists A : (\forall x : x \in A \leftrightarrow \phi))$ for each stratified ϕ in which A is not free.

Quine's original definition of stratification is quite elegant, but we will not give it, as it would take us into a technical morass.

The usual modern definition is that we say ϕ is stratified iff there is a function τ from the variables of ϕ to natural numbers such that for each atomic subformula $u = v$ of ϕ we have $\tau(u) = \tau(v)$ and for each atomic formula $u \in v$ of ϕ we have $\tau(u) + 1 = \tau(v)$. The connection of this to well-formedness conditions of formulas in TST should be clear.

There is an accusation that the stratification criterion is a mere syntactical trick. This is refutable. NF can be formulated without any reference to types at all, because the stratified comprehension axiom is equivalent to the conjunction of finitely many of its instances.

Quine made at least two serious mistakes in the 1937 paper.

One, which had no lasting consequences but is an embarrassing slip on the part of someone who knew better, is the claim that NF proves Infinity because any model contains infinitely many objects, the empty set and its iterated singletons. This only shows that all models of the theory are infinite. In fact, the version NFU in which extensionality is weakened to allow atoms of course has only infinite models for the same reason, but is consistent with the negation of the Axiom of Infinity. NF *does* prove Infinity but for very different and rather alarming reasons.

The second, which is an essential error which led to a vast amount of trouble and effort for mathematicians (including our need to prepare this talk) was the choice of strong extensionality as the extensionality axiom of NF. Quine considers the alternative axiom

$$(\forall xyz : z \in x \land (\forall w : w \in x \leftrightarrow w \in y) \rightarrow x = y)$$

which weakens extensionality to allow many objects with empty extension. Quine believed that he could easily eliminate atoms if he allowed them, so that he could harmlessly assume strong extensionality. He was wrong, and we may discuss the nature of the error presently.[*] As we will explain below, Jensen showed in 1969 that the theory NFU with the weaker extensionality axiom is quite readily shown to be consistent.

[*] in the Q & A period, it did not make it onto the slides.

The disaster of 1953

In 1953, Ernst Specker proved that NF disproves the Axiom of Choice. This was a complete surprise to all concerned. A corollary is that NF proves the Axiom of Infinity.

Also in 1953, Rosser's lovely book *Logic for Mathematicians* appeared. This book used NF as the basis for a presentation of the foundations of mathematics. It was not explicitly discredited by Specker's bombshell result, because Rosser did not assume Choice (although he proved equivalence of Choice with various statements); he prudently adopted Countable Choice as an axiom, and we can report that the system of Rosser's book is actually consistent. He also introduced notational refinements which subsequent workers in NF have found useful, and proposed another important axiom extending the theory, his Axiom of Counting, which we may or may not at some point discuss. It is a beautiful book, but its fate was clouded by Specker's result.

Specker justifies the motivation behind Quine's bold move

In 1962, Specker proved the following interesting result. The following theories are equivalent in consistency strength:

- NF
- TST + the Ambiguity Scheme, the collection of axioms $\phi \leftrightarrow \phi^+$ (ϕ closed)
- TST + a type shifting endomorphism (an external map σ such that $\sigma(x)$ is one type higher than σ, $(\forall y : \exists x : \sigma(x) = y)$, $\sigma(x) \in \sigma(y) \leftrightarrow x \in y$ and $\sigma(x) = \sigma(y) \leftrightarrow x = y$.) Being "external", σ cannot be used in instances of comprehension. This is usually expressed as "there is a model of TST with a type shifting endomorphism", but there is a theory of such models.

The effect of this was to justify Quine's bold move formally. It didnt show that NF is consistent: it showed that Quine's move from noticing typical ambiguity to actually identifying the types made sense.

Enter NFU

At this point, everyone (or everyone who was interested) was anxious about whether NF made sense at all. Might it be inconsistent?

In 1969, Jensen showed that stratified comprehension is consistent. More precisely, he showed that the theory NFU which results if one replaces the axiom of extensionality with the weaker form which allows atoms is consistent. Further he showed that it is consistent with Infinity, consistent with Infinity and Choice and has ω-models. He also showed that it is consistent with the negation of Infinity!

NFU + Infinity + Choice is as competent for mathematical work as TST + Infinity + Choice, which supports all of classical mathematics outside of set theory, and without the weirdness of the hierarchy of types and the hall of mirrors.

This result justified Quine's 1937 set theoretical program (if he actually had such a program, which is not altogether clear). Most people didn't notice. I'm not saying much of anything about the paradoxes of set theory here, but I don't need to: the usual paradoxes have solutions in NFU, which we could examine in actual models since 1969, which creates confidence that they are not problems for NF.

A small technical point: Jensen did not do this, but NFU can be enhanced with a constant for the empty *set* (as distinct from the many possible atoms) in each type, and this makes translation of mathematical work into NFU less laborious.

It is useful to sketch a proof of Jensen's result. Let λ be a limit ordinal. Observe that each strictly increasing sequence s in λ can be associated with a model of TSTU, in which type i is implemented as V_{s_i}, equality in each type is equality restricted to that type, and membership of type i objects in type $i+1$ objects is defined by $x \in_i y$ iff $x \in V_{s_i} \wedge y \in V_{s_i+1} \wedge x \in y$. Note that each element of $V_{s_{i+1}} - V_{s_i+1}$ is construed as an urelement (in its capacity as a type $i+1$ object; as an object of any higher type it has its usual extension; the empty set in V_{s_i+1} is the designated empty set in type $i+1$; it is an odd feature than the types of this model are nested rather then disjoint, but it works perfectly well, because TST just does not address the question whether objects of different types are equal or not).

Let Σ be any finite set of formulas in the language of TSTU. Let $n-1$ be the highest type appearing in any formula in Σ. The formulas in Σ determine a partition of $[\lambda]^n$: an n-element subset A of λ is put in a compartment determined by the truth values of the formulas in Σ in models of TSTU determined as above in which $s``\{0,\ldots,n-1\} = A$. Now this partition has a homogeneous set by Ramsey's theorem. Let h be an increasing sequence in the homogenous set. The model determined by h as above will satisfy $\phi \leftrightarrow \phi^+$ for each $\phi \in \Sigma$, by homogeneity. So each finite subset of the Ambiguity Scheme is consistent with TSTU, so TSTU + ambiguity is consistent by compactness. Specker's ambiguity results of 1962 also hold for TSTU and NFU, so NFU is consistent.

Note that we can require Infinity to hold in the model of TSTU associated with any sequence by providing that $\lambda > \omega$ and requiring that all ordinals s_i be infinite: this gives a proof of the consistency of TSTU + Infinity + Ambiguity. We can require the negation of Infinity to hold in the model of TSTU associated with any sequence by letting $\lambda = \omega$. This gives a proof of the consistency of TSTU+Ambiguity + the universe is finite. If Choice holds in the metatheory, Choice can be assumed as part of the base theory holding in every sequence.

So we have consistency of NFU + Infinity + Choice and of NFU + the negation of Infinity, by Specker's 1962 results. The Specker proofs of the negation of Choice and so of Infinity do not go through in NFU.

Another approach

A closely related approach attributed to Boffa is to employ a model of ordinary set theory in which there is an external automorphism moving a rank of the cumulative hierarchy, wlog downward.

Let j be the external automorphism. Let V_α be the (necessarily nonstandard) level moved. The domain of the model of NFU to be constructed is V_α. Equality of model elements is interpreted by true equality. $x \in_{\mathtt{NFU}} y$ is defined as $j(x) \in y \wedge y \in V_{j(\alpha)+1}$. Notice that $V_{j(\alpha)+1}$ is an isomorphic copy of $V_{\alpha+1}$, which the nonstandard model thinks is the power set of our universe V_α, and all elements of $V_\alpha - V_{j(\alpha)+1}$ are interpreted as urelements.

We will not give the proof that this is a model of NFU. But we regard this construction as valuable for getting a mental picture of what the world of NFU is like.

Quine's set theoretical program (if he had one) was at this point justified.

NFU + Infinity + Choice is a fully competent set theory, usable for foundational purposes. It has its advantages and disadvantages, like any foundational system. I would say that the choice of Zermelo set theory and its extensions which the mathematical community has made historically is a better one, but one can imagine a world in which NFU was adopted as the consensus scheme.

NFU + Infinity + Choice (as I have explained at length elsewhere) can be extended with strong axioms of infinity much as ZFC can, to give far stronger systems.

It is a lovely subject, but not part of the aim of this paper, to give an account of mathematics in stratified set theory...we will avoid doing this here, and proceed along our main track, the story of NF itself.

NF was proposed in 1937. Specker cast it into doubt by demonstrating the failure of Choice in 1953. Specker demonstrated that Quine's collapse of the types on the basis of ambiguity made sense in 1962. Jensen proved the consistency of NFU in 1969. We fast forward to Holmes (the speaker blushes slightly) 1995.

Jensen's proof depends on the fact that any sequence of V_α's determines a model of TSTU: we can, for $\alpha > \beta$ construe V_α as the power set of V_β ($V_{\beta+1}$ plus additional elements (the elements of $V_\alpha - V_{\beta+1}$) which we construe as urelements. This allows us to apply Ramsey's theorem to get sequences which are uniform in their treatment of (a finite set of) statements in the language of TSTU.

Of course you cannot do this with actual power sets. But you can imagine faking it.

In my paper of 1995 I proposed a type theory TTT (tangled type theory). This is a first order theory with sorts indexed by the natural numbers (or by any limit ordinal λ) with equality and membership as primitive predicates. $x = y$ is well-formed iff the types of x and y are the same. $x \in y$ is well-formed iff the type of y is greater than the type of x.

If ϕ is a formula in the language of TST and s is a strictly increasing sequence of type indices, and we have a bijection from type i variables of TST to type j variables of TTT whenever $i \leq j$, we define ϕ^s as the result of replacing each variable of type i in ϕ with the variable of type s_i computed using the appropriate bijection. ϕ^s will clearly be well-formed. The axioms of TTT are exactly the formulas ϕ^s where ϕ is an axiom of TST.

This theory is extremely strange. To begin with, it is *not* cumulative type theory. Each object of type α is completely determined by its type β elements, and also completely determined by its type γ elements, where $\beta \neq \gamma$ are ordinals less than α. Each type α contains all collections of type β objects which can be defined as in TST using types restricted to a specific sequence.

Each type α is a kind of power set of each type $\beta < \alpha$. Cantor's theorem in the metatheory shows us that most of these power sets cannot be honest. In fact, no type in the structure can be any larger than the true power set of type 0.

This theory is equiconsistent with NF. One direction is easy: a model of NF can be used to implement every type of a model of TTT. If $(M, \in_M)$ is a model of NF, implement each type α as $M \times \{\alpha\}$ and implement membership of the model N of TTT as
$(m, \beta) \in_N (n, \alpha) \equiv_{\texttt{def}} m \in_M \wedge \beta < \alpha$.

The other direction is a recapitulation of Jensen's proof of the consistency of NFU.

Suppose we have a model of TTT. Let Σ be any finite set of formulas in the language of TST. Let $n-1$ be the highest type appearing in any formula in Σ. The formulas in Σ determine a partition of $[\lambda]^n$: an n-element subset A of λ is put in a compartment determined by the truth values of the formulas ϕ^s for $\phi \in \Sigma$ in the model of TTT determined by strictly increasing sequences s for which $s``\{0, \ldots, n-1\} = A$. Now this partition has a homogeneous set by Ramsey's theorem. Let h be an increasing sequence in the homogenous set. The model of TST determined by h, in which type i is implemented as type s_i of the model of TTT, will satisfy $\phi \leftrightarrow \phi^+$ for each $\phi \in \Sigma$, by homogeneity. So each finite subset of the Ambiguity Scheme is consistent with TST. Specker's ambiguity results of 1962 show that NF is consistent.

The problem of establishing the consistency of NF "reduces" to the problem of finding a model of TTT. A model of TTT has the positive merit of being a well-founded structure – if there is one. But TTT is so extraordinarily weird that my reaction to seeing that this was equivalent to NF was to abandon any belief that NF was consistent (or belief that it was not).

At the same time, I gave a definition of an extension of Mac Lane set theory (bounded Zermelo without choice) containing a strange system of cardinals called a "tangled web": existence of a model of Mac Lane with a tangled web entails consistency of NF, again by an adaptation of Jensen's proof. I give the definition of a tangled web but not the proof.

A tangled web of cardinals is a function τ from finite subsets of λ (a limit ordinal) to cardinals. satisfying conditions (1) and (2) stated below. In the absence of choice, one can use the Scott definition of cardinal.

For A a finite subset of λ, define A_1 as $A \setminus \{\mathtt{min}(A)\}$.

(1) $\tau(A_1) = 2^{\tau(A)}$ for each A.

(2) The first order theory of the natural model of TST$_n$ (TST using only n types) in which type 0 is a set of cardinality $\tau(A)$ and type $i+1$ is the power set of type i is determined exactly by the n smallest elements of A.

We do not give the proof that existence of a tangled web entails Con(NF) [it has similar flavor to the Jensen proof of consistency of NFU]. The interest of this is again that one is considering well-founded and probably not even very large structures. But again, it is very hard to see why the situation described should even be possible.

A purely technical side remark: existence of an actual tangled web as described is (we think) stronger than Con(NF). NF is equivalent to something like existence of tangled webs of arbitrary concrete finite size (A being a subset of a natural number, not a limit ordinal).

From 1995 to 2010 I was doing other things, and this was percolating in my mind.

From 2010 onward I began to claim that I had a proof of the existence of a model of ZF in which there was a tangled web of cardinals (which would entail the consistency of NF). The “proof” was insanely complicated and hard to read, and early versions certainly contained errors. It is a project of mine to actually write out a full version of this argument and see if I can make it clear that it works (or that it doesn’t). But it is definitely the ancestor of the (still insanely complicated and hard to read, but evidently correct) proof of the existence of a model of TTT which I have more recently produced, which was verified in the Lean prover by Sky Wilshaw in 2024.

The idea which made the model construction possible was a concrete description of the alternative extensions of each object in the model of TTT. Each set which we define by an abstract $\{x^{s_i} : \phi^s\}^{s_{i+1}}$ has an intended extension in type s_i: what are all of its other extensions? The brief answer is that we provide ourselves with a suitable collection of highly symmetrical junk from which to build all the alternative extensions. Details follow. In addition to the types indexed by ordinals $< \lambda$ which will make up our model of TTT, we provide a type -1 made up of μ atoms, where μ is the common cardinality of all the types. We specify a regular uncountable cardinal κ (and define "small" as "is of cardinality $< \kappa$) and a partition of type -1 into sets of size κ called "litters". Subsets of type -1 with symmetric difference of size $< \kappa$ from a litter are called "near-litters". We will arrange that the model will see the litters as κ-amorphous: any subset of a litter is either small or a near-litter, as far as the model is concerned.

We give some general scaffolding for the model as a structure in the metatheory. Type -1 is a set of size μ given in advance. Each type $\alpha > -1$ has each of its inhabitants a collection of elements of the union of types $\beta < \alpha$ plus a type label (which is not a model element). No type -1 object contains any of the type labels; no object contains more than one type label. This is enough for the types to be disjoint. Of course, not all sets meeting these specifications can be elements of the model.

We use the notation τ_α for the set implementing type α. For any $\beta < \alpha$, we can conveniently write $x \cap \tau_\beta$ for the extension of x in type β: the membership relation of our eventual model of TTT is a subset of the membership relation of the metatheory.

We associate sets of litters X_α of size μ with each type index in $\{-1\}\cup\lambda$. We choose objects in each type α with -1-extension each near-litter. We note, aside, that type -1 will not be part of the model of TTT, and the higher types satisfy weak extensionality over it: many objects in each type have no type -1 elements.

We can then describe the bare bones of the scheme of extensions. Each object x of type α has a distinguished extension in a specific type $\beta < \alpha$ We can write this $x \cap \tau_\beta$. There is a uniform way to compute $x \cap \tau_\gamma$ for every other $\gamma < \alpha$. If $\beta \neq -1$, $x \cap \tau_{-1}$ is empty. Suppose $\beta > -1$. We provide injective maps $f_{\beta,\gamma}$ from type β to X_γ. All distinct f maps have disjoint ranges. $x \cap \tau_\gamma$ is then the collection of all typed near-litters N_γ such that N has small symmetric difference from an element of the range of $f_{\beta,\gamma}``(x \cap \tau_\beta)$. N_γ is the type γ model element whose intersection with τ_{-1} is N. Notice that our scheme actually indicates what the rest of its elements are.

There are further rules about distinguished extensions. A model element with $x \cap \tau_{-1}$ nonempty has that as its distinguished extension. A model with empty extension over any type other than type -1 has its empty extension over type -1 as its distinguished extension. Obviously a model element with an extension which is not a collection of near-litters closed under small symmetric difference containing litters from a single X set has that as its distinguished extension (since all but one of the extensions of any object must be of this form).

Note that if we are given a set A which can be a non-distinguished extension, we can uniquely compute the distinguished extension from which it would have been derived: it will be a collection of typed near-litters with extensions from a particular X_β which one can then determine a subset of τ_β. We denote this hypothetical distinguished extension by $\delta(A)$. We stipulate (we can arrange this by clever construction of the f maps) that no candidate extension for an object in the model has infinitely many iterated images under δ: then we can require that distinguished extensions have an even number of iterated images under δ.

Notice that when we are considering the extensions of a type α object, we need only consider $f_{\beta,\gamma}$'s with indices $< \alpha$.

The conditions described here ensure that the model structure, whatever it is, is extensional over types other than type -1: as soon as you are given a distinguished extension of a model element, you can compute all the others, and no extension can be a distinguished extension of one model element and a non-distinguished extension of another.

It remains to decide which sets actually belong to the model, and in such a way that TTT comprehension is satisfied. The idea is that the sets of the model are those which are symmetric under a family of permutations which is determined precisely by the scheme for defining alternative extensions given above.

A structural permutation of type α is a permutation π of type α such that there is a permutation π_β of each type $\beta < \alpha$ such that for each $x \in \tau_\alpha$, $\pi(x) \cap \tau_\beta = \pi_\beta``(x \cap \tau_\beta)$. Note that if all types below α have been defined, and we are given permutations π_β for each $\beta < \alpha$, we can compute π on all subsets which could be model elements. [this is not the definition of structural permutation in the paper; it is a convenience for these slides].

A structural permutation thus defined must satisfy a coherence condition. For each $y \in \tau_\beta$, we expect there to be an $x \in \tau_\alpha$ such that $x \cap \tau_\beta = \{y\}$. This must be the distinguished extension of x. Thus for each $\gamma \in \alpha - \{-1, \beta\}$, $x \cap \tau_\gamma$ is the set of near-litters with small symmetric difference from $f_{\beta,\gamma}(y)$. Now π maps x to the unique z such that $x \cap \tau_\beta$ is $\{\pi_\beta(y)\}$, and $z \cap \tau_\gamma$ is the set of typed near-litters with small symmetric difference from $f_{\beta,\gamma}(\pi_\beta(y))$, but it is also the set of all $\pi_\gamma(N_\gamma)$, where N is a near-litter with small symmetric difference from $f_{\beta,\gamma}(y)$.

So the coherence condition that

$$\{\pi_\gamma(N_\gamma) : N \sim f_{\beta,\gamma}(y)\} = \{N_\gamma : N \sim f_{\beta,\gamma}(\pi_\beta(y))\}$$

must hold for any structural permutation.

We then define an allowable permutation as a structural permutation π for which all the maps π_β are allowable.

Notice that this allows a refinement of the coherence condition: π_γ is seen to be allowable and so structural, so it is sufficient to require that $(\pi_\gamma)_{-1}``f_{\beta,\gamma}(y) \sim f_{\beta,\gamma}(\pi_\beta(y))$ as the coherence condition when only allowable permutations are considered.

The executive summary of what follows is that the sets of the model are those that are fixed by all allowable permutations that fix suitable supports. There are complexities because the appropriate notion of support is quite complicated.

Where A is a nonempty finite subset of $\lambda \cup \{-1\}$ and π is an allowable permutation on type α, we define $\pi_{\{\alpha\}}$ as π and define $\pi_{A \cup \{\beta\}}$ as $(\pi_A)_\beta$ exactly when this makes sense (when $\beta < \mathtt{min}(A)$).

A support condition is a pair (x, A) where x is a near-litter or a type -1 object and A is a finite subset of $\lambda \cup \{-1\}$ containing -1.

A support is a small set of support conditions in which all second projections have the same largest element (if we call this α we call it an α support).

The action of an α support on a support S is defined by $\pi[S] = \{(\pi_A``x, A) : (x, A) \in S\}$.

An element x of type α has support S iff S is an α support and all allowable permutations π such that $\pi[S] = S$ also satisfy $\pi(x) = x$.

The elements of type α are exactly the candidate elements of type α which have supports. Since we can define allowable permutations on all candidate elements before actually determining what type α's extent is, this definition works. A candidate element of type α is determined by a possible distinguished extension in some lower type, and we have (almost, mod fiddly details) described exactly what the possible distinguished extensions are for type α objects with no knowledge of the extent of type α.

Once type α is constructed, we need to build maps $f_{\alpha,\beta}$ for each type $\beta > -1$ [we can do this for all higher types as well].

We have given an almost complete description of an early description of the construction of the model (we left out some fiddly stuff about choosing the f maps. To avoid being disingenuous, we must comment that the definition of the f maps have been complicated so that one maps not bare model elements in one type to litters in another type, but model elements equipped with supports to litters in another type. The older description gives an adequate impression of the idea of the proof: the extra bells and whistles make the proof easier to write and verify.

The proof that this works is elaborate, and we give simply the highlights.

It is necessary to prove that each type α is of size μ. The requirements on μ are that it be a strong limit cardinal $\geq \lambda, \kappa$ with cofinality at least κ. This is the really hard part of the proof.

It is necessary to prove that predicative comprehension holds (that each axiom $\{x^{s_i} : \phi^s\}^{s_{i+1}}$ holds, where no variable of type higher than s_{i+1} appears, and no bound variable of type s_{i+1} appears). This is actually fairly easy!

To get full comprehension, one has to show that the many typed versions of the axiom of union hold in addition, meaning that one can lift the predicativity restriction on comprehension. This is not terribly hard to prove, but it is surprising, perhaps, that it works.

Although I do not have to say this in presenting the argument, it is basically a Frankel-Mostowski construction (perhaps the most complicated one ever done?) The original plan of building a tangled web in a model of ZFC was explicitly an FM construction.

In some ways, this is a boring resolution of the NF consistency problem. For every type β and each $\alpha > \beta$, and each small subset A of τ_β, there is an element $A_\alpha \in \tau_\alpha$ such that $A_\alpha \cap \tau_\beta = A$. This can briefly be expressed by saying that the model is κ complete. So for familiar mathematical structures of bounded size, choose κ large enough and our model of TTT will see these structures exactly as the metatheory sees them.

TTT (and so NF) has nothing new to reveal about the Continuum Hypothesis for example, and it is consistent with Countable Choice (the axiom in Rosser's book) and with Dependent Choices. It should be noted that models of NF are *not* κ complete (they cannot even be ω-complete) but a model of NF obtained from a model of TTT will have the same theorems.

A specific startling result which is not of this local nature is that our models satisfy the principle that the power set of a well-ordered set is well-ordered. This is enough choice for all applications in analysis, for example, and suggests that NF (like NFU) actually supports mathematics in a sensible if somewhat unfamiliar style.

Open questions remain: for example, can the universe be linearly ordered in NF? Our construction excludes this, because of the role played by κ-amorphous sets, but this isn't clearly essential. Our methods give no clue as to what general models of NF look like: I made very specific and concrete assumptions about alternative extensions to build this model, which have strong consequences which are not necessarily consequences of NF per se.

We don't know what the strength of NF is. We believe the consistency strength of NF to be exactly that of TST + Infinity. However, the minimal values of the parameters in our proof are $\lambda = \omega, \kappa = \omega_1, \mu = \beth_{\omega_1}$. Thus we need existence of all countable ordinal indexed ranks of the cumulative hierarchy to support our proof, at least. We believe this can be sharpened, but we do not yet know how to do this.

Note that the Lean formalization proves that there is a model of TTT. The arguments from a model of TTT to consistency of NF have not been formalized, but they are off the shelf variants of the argument of Jensen. There is a plan to produce a formal proof of the actual statement Con(NF) for our own satisfaction.

Lean itself formally requires lots of indiscernibles, but it should be evident that no essential use of these is made in our argument.