An outline of a proof of the consistency of New Foundations

M. Randall Holmes
Boise State University

June 11, 2021

The Problem

In 1937, W. v. O. Quine, a notable Americal philosopher and logician, proposed what is perhaps the most streamlined possible version of Russell's theory of types from Principia Mathematica, in a paper titled New Foundations for Mathematical Logic, from which the theory is usually called NF (for "New Foundations").

New Foundations is a variation on TST, the simple typed theory of sets. This is the sorted theory with equality and membership as primitive predicates, the sorts being indexed by natural numbers and the legal forms for atomic sentences neatly summarized by the schemata $x^i = y^i$, $x^i \in y^{i+1}$, and axioms of extensionality and comprehension exactly as in naive set theory (comprehension being restricted by the rules for formation of sentences: $\{x : x \notin x\}$ is not provided by an instance of comprehension because a sentence of the shape $x \in x$ is not well-formed, no matter what the type of x is).

It is an interesting historical note that TST is **not** the type theory of Principia Mathematica, though a summary of TST is often presented as an account of PM by careless writers. TST appears to be described first by Tarski around 1930. The obstruction to Russell and Whitehead presenting their theory in this way is that they had no idea how to implement the ordered pair using sets, so the type system of PM is a complicated system of relation types, further complicated by predicativity considerations. Norbert Wiener appeared to have TST in mind when he presented the first set theoretical definition of the pair in 1914, but he did not give a formal description.

NF is motivated by observing the phenomenon of systematic ambiguity in TST, which Russell had already noted in the system of PM. This symmetry is much more striking in TST. Provide an injective operation $x \mapsto x^+$ on variables which raises type by one. Let ϕ^+ be the result of replacing each variable x in ϕ with x^+ throughout. Then ϕ^+ is a theorem of TST if ϕ is a theorem (the converse is not true: more can be proven about higher types), and any object defined in set builder notation in a form $\{x : \phi\}$ has an exact analogue $\{x^+ : \phi^+\}$ in the next higher type (and this can be iterated).

Quine's proposal was that it seems reasonable with such a high degree of symmetry to suppose that the types are simply **the same**: the resulting theory is a single sorted theory with equality and membership, with axioms of extensionality and a comprehension scheme consisting of those assertions "$\{x : \phi\}$ exists" which can be obtained from instances of the comprehension scheme of TST by ignoring distinctions of type between the variables.

It is traditional to give an account of the comprehension scheme of NF which does not mention the rules of sentence formation of another theory. A function σ from variables to natural numbers is called a stratification of a formula ϕ if each subformula '$x = y$' of ϕ satisfies σ('x') $=$ σ('y') and each subformula '$x \in y$' of ϕ satisfies σ('x') $+ 1 =$ σ('y'). A formula ϕ is said to be stratified iff there is a stratification of ϕ. The comprehension axiom of NF can be presented in the form "$\{x : \phi\}$ exists if ϕ is stratified". This has been criticized as a syntactical trick. It is worth noting that stratified comprehension is equivalent to a finite subset of its instances, so in fact it can be expressed in a way which makes no reference to types at all. A finite axiomatization can be developed by analogy with the finite axiomatization of the class comprehension axiom scheme of von Neumann-Gödel-Bernays class theory.

We will not allow ourselves to be distracted too much about the oddities of the way the world looks in this theory. The universal set exists and sets make up a Boolean algebra under the usual operations. The Frege natural numbers exist and are the natural implementation of $\mathbb{N}$. More generally, Russell-Whitehead cardinals and ordinals exist and are the natural implementations of cardinal and ordinal numbers. The Russell paradox is trivially avoided ($x \not\in x$ is not a stratified formula). The Cantor paradox is avoided because the form of Cantor's theorem in TST is $|\iota``A| < |\mathcal{P}(A)|$ $[\iota = (x \mapsto \{x\})]$ instead of the ill-typed $|A| < |\mathcal{P}(A)|$. $|A| < |\mathcal{P}(A)|$ is a well-formed assertion in the language of NF, but clearly false if $A = V$, the universal set. $|\iota``V| < |\mathcal{P}(V)|$ is provable, so the singleton operation is not implemented by a set function: the collection of singletons is smaller than the universe in spite of the obvious external bijection between these sets. The way in which NF avoids the Burali-Forti paradox is fascinating but would take us too far afield.

The first indication that NF is not a harmless notational variation of TST was Specker's 1954 result that the Axiom of Choice is refutable in NF. We won't discuss the details of the proof of this result at all, but the bare fact may serve to provide a hint of the motivation behind the approach we took to the problem.

At this point if not before it was clear that there is a **problem** of consistency of NF relative to a set theory in which we have confidence. That NF proves the negation of Choice shows that NF is at least as strong as TST + Infinity: there is no evidence that NF is any stronger than TST + Infinity, which has the same strength as Zermelo with Separation restricted to bounded formulas.

In 1962, Specker proved that NF is equiconsistent with TST with the Ambiguity Scheme asserting $\phi \leftrightarrow \phi^+$ for each closed formula ϕ. This result is not trivial to prove, though it is strongly suggested by Quine's original motivation in defining this theory.

In 1969, R. B. Jensen proved that NFU (New Foundations with urelements), the theory obtained by modifying NF by weakening extensionality to apply only to objects with elements, is consistent and moreover is consistent with Infinity and with Choice (it is consistent with stronger axioms of infinity as well, and this is clear from Jensen's original paper).

Specker's result can be generalized to show that NFU is equiconsistent with TSTU + Ambguity, where TSTU has extensionality weakened to allow many objects with no elements in each positive type.

Jensen's consistency proof shows that the ways that NF avoids the Russell, Cantor, and Burali-Forti paradoxes work in a consistent set theory. If NF were inconsistent, it would fall prey to a different paradox. Specker's proof can be understood as showing that NF + Choice falls victim to a different paradox of set theory.

It is another interesting matter of history to note that Quine discusses his choice of strong extensionality as an axiom for NF: he claims it can be harmlessly motivated by the identification of urelements u with their singletons $\{u\}$, which is an actual mathematical error in a context with stratified comprehension. It would take us too far afield to discuss the details, but it is worth noting for the benefit of anyone who might remember this from looking at the original paper.

The second step to the solution: tangled type theory and tangled webs (Holmes 1995)

The first step to the solution is Jensen's 1969 proof of the consistency of NFU. I'll be able to retrospectively give an account of Jensen's proof after I give an account of my 1995 results.

Tangled type theory

TTT (tangled type theory) is a theory with sorts indexed by ordinals less than a limit ordinal λ (which may be taken as ω but does not have to be) with formation rules for atomic sentences $x^\alpha = y^\alpha$ and $x^\alpha \in y^\beta$ for $\alpha < \beta$.

For any finite sequence s of ordinals less than λ and formula ϕ in the language of TST, define ϕ^s as the formula in the language of TTT obtained from ϕ by replacing each variable x^i in ϕ with x^{s_i}. (We can make exact sense of this by requiring in both theories that each variable is of the form $\mathbf{x}_n^\tau$ where n is a natural number index and τ is a type).

The axioms of TTT are the formulas ϕ^s where ϕ is an axiom of TST and s is any strictly increasing sequence of ordinals less than λ.

For mental hygiene, I strongly suggest not thinking about what the world looks like in TTT. Focus simply on the formal proof which follows that TTT is exactly as strong as NF.

From a model of NF one immediately obtains a model of TTT by using a copy of the model of NF to implement each type in the model of TTT, the membership relations between copies being determined by the membership relation of the model itself in the obvious way.

Suppose that we are given a model of TTT.

Let Σ be a finite set of sentences in the language of TST. Let n be a strict upper bound on types mentioned in formulas in Σ. We use Σ to define a partition of the collection $[\lambda]^n$ of n element subsets of λ: the compartment in which $A \in [\lambda]^n$ is put is determined by the truth values of formulas ϕ^s for $\phi \in \Sigma$ and $\mathtt{rng}(s\lceil n) = A$. This partition has an infinite homogeneous set H which includes the range of a strictly increasing sequence h. The model of TST determined in the obvious way by the sequence h of types taken from λ (ϕ holds in this model iff ϕ^h holds in our model of TTT) satisfies TST $+$ (Ambiguity restricted to formulas in Σ). But this implies directly that TST $+$ Ambiguity is consistent by compactness, so NF is consistent by the results of Specker.

This argument is motivated by Jensen's original proof of Con(NFU) and can be reverse engineered to a proof of Con(NFU). The idea is that TTTU has an obvious model: let type α $(\alpha < \lambda)$ be implemented by V_α, the level of the cumulative hierarchy indexed by α, with $x^\alpha \in y^\beta$ taken as meaning $x^\alpha \in y^\beta \wedge y^\beta \in V_{\alpha+1}$, where x^α ranges over V_α and y^β ranges over V_β in the familiar set theoretical universe. Notice that all elements of $V_\beta \backslash V_{\alpha+1}$ are being treated as urelements here. This can further be seen to show that NFU has models exhibiting stratified versions of whatever mathematical statements we might expect to be true in a suitably chosen level of the cumulative hierarchy: NFU is only superficially different from ordinary set theory, in a sense which we have discussed at length elsewhere.

TTTU has natural models which we have just described. Natural models of TTT are quite another matter. The difficulty is that each type is being interpreted as a "power set" of *each* lower type, not just an immediate predecessor, and simple considerations of cardinality make it clear that something very strange has to be going on for this to be achieved.

By contrast, in the argument for Con(NFU) each type is being interpreted as the power set of each lower type...plus lots of urelements in each case. And we have indicated how to do this straightforwardly.

We modify (as it were "unfold") this picture to something which may appear achievable if still weird in the ordinary set theoretical universe.

Oh what a tangled web we weave...

We work in ordinary set theory without choice and with extensionality weakened to allow a set of atoms, in which we can use the Scott definition of cardinal (this requires nothing more than bounded Zermelo set theory plus the condition that every set belongs to a rank of the cumulative hierarchy (level 0 of the hierarchy being the set of atoms)).

Let λ be a limit ordinal. For any nonempty finite subset A of λ, define A_1 as $A \setminus \{\mathtt{min}(A)\}$. Define A_0 as A and A_{n+1} as $(A_n)_1$.

TST$_n$ is the subtheory of TST in which only types less than n are used (so there are n types, the lowest one being 0). A natural model of TST$_n$ is one in which the lowest type is a set X, each type $i < n$ is implemented by $\mathcal{P}^i(X)$, equality within each type is the equality of the ambient set theory, and membership of objects of each type in objects of the next type is the membership relation of the ambient set theory. Esthetically one might prefer that the types be disjoint and this can be arranged but is not actually needed.

It is an important fact that the theory of a natural model of TST$_n$ depends only on n and the cardinality of the set implementing type 0.

A *tangled web* is a function τ from nonempty finite subsets of λ to cardinals with the following properties:

naturality: for any A with $|A| > 1$, $2^{\tau(A)} = \tau(A_1)$

elementarity: for any A with $|A| > n$, the theory of a natural model of TST$_n$ with type 0 implemented by a set of size $\tau(A)$ depends only on $A \setminus A_n$ (the set consisting of the n smallest elements of A).

The existence of a tangled web implies the consistency of NF. The proof is very similar to the proof given above for tangled type theory.

Let Σ be a finite set of sentences in the language of TST. Let n be a strict upper bound on types mentioned in formulas in Σ. We use Σ to define a partition of the collection $[\lambda]^n$ of n element subsets of λ: the compartment in which $A \in [\lambda]^n$ is put is determined by the truth values of formulas $\phi \in \Sigma$ in natural models of TST$_n$ with base types of size $\tau(B)$ where $B \setminus B_n = A \neq B$. This partition has a homogeneous set B of size $n+2$.

We observe that a natural model of TST_{n+1} with base type of size $\tau(B)$ has the same truth values for Σ as the model of TST_n whose base type is the set implementing type 1 in the previous model whose size is $2^{\tau(B)} = \tau(B_1)$, by homogeneity of B with respect to the indicated partition. The punchline is that that TST_{n+1} is consistent with ambiguity for sentences in Σ, so TST is consistent with ambiguity for sentences in Σ, so TST + Ambiguity is consistent by compactness.

It is worth noting somewhere, and it might as well be here, than both TTT and ordinary set theory minus Choice with a tangled web disprove Choice, in a manner which can be developed by analogy with Specker's disproof of Choice in NF.

The final move: constructing a tangled web

There is a version of the proof which involves constructing a model of tangled type theory directly. Tangled type theory is bewildering; the earliest versions of the argument took the approach of constructing a tangled web, and that is what we will do here.

The argument is via construction of a Frankel-Mostowski permutation model in ZFA. A very simple argument of this kind was originally presented to show that Choice is independent of ZFA.

There are some cardinal invariants of the construction.

λ is a limit ordinal.

For finite subsets A of λ with at least n elements, we define A_n as above.

κ is a regular uncountable cardinal. A set of size $< \kappa$ is called small; all other sets are called large.

μ is a strong limit cardinal of cofinality greater than λ or κ.

We work in ZFA with μ atoms.

We specify an order $\ll$ on finite subsets of λ. It is uniquely specified by three conditions:

If $A \neq \emptyset$, then $A \ll \emptyset$

If $\mathtt{max}(A) < \mathtt{max}(B)$, then $A \ll B$

If $\mathtt{max}(A) = \mathtt{max}(B)$, then $A \ll B$ iff $A \backslash \{\mathtt{max}(A)\} \ll B \setminus \{\mathtt{max}(B)\}$.

Note that $A \ll A_i$ if $i > 0$: downward extensions of a set appear before the set in this order.

There are μ atoms, partititioned into sets τ_A^0 for each finite subset A of λ; each of these sets is of size μ.

Each set τ_A^0 is partitioned into sets of size κ called *litters*. A subset of a τ_A^0 with small symmetric difference from a litter is called a *near-litter*. For any litter L, the set of all near-litters with small symmetric difference from L is called the local cardinal of L, written $[L]$.

We define τ_A^1 as the collection of all subsets X of τ_A^0 for which there is a small set Y of litters included in τ_A^0 such that either $X \Delta \bigcup Y$ is small or $X \Delta (\tau_A^0 \setminus \bigcup Y)$ is small: that is, X has small symmetric difference from a small or co-small union of litters included in τ_A^0.

For each α we choose a map χ_α which is an injective map with domain the union of all τ_A^0 with $\mathtt{max}(A) = \alpha$ whose restriction to each such τ_A^0 is a bijection from τ_A^0 to $\tau^0_{A \setminus \{\mathtt{max}(A)\}}$. We further require that the elementwise image of a litter under χ_α is a litter.

We extend the action of χ_α to any set whose transitive closure contains no atoms not in its domain by the rule $\chi_\alpha(X) = \chi_\alpha``X$.

We will construct for each A a set $\tau^2_A \subseteq \mathcal{P}(\tau^1_A)$. All these sets are of cardinality μ. [In the eventual FM model, the cardinality of τ^2_A will be $\tau(A)$, where τ is the desired tangled web.]

We define K_A as the collection of local cardinals of litters included in τ^0_A. We will provide for each singleton $\{\alpha\}$ a map $\Pi_{\{\alpha\}}$, a bijection from $K_{\{\alpha\}}$ to the union of $\tau^0_\emptyset$ and all sets $\tau^2_{\{\alpha,\beta\}}$ for which $\beta < \alpha$. For each A with $|A| > 2$ we define Π_A as $\chi_{\mathtt{max}(A)}(\Pi_{A \setminus \{\mathtt{max}(A)\}})$. It follows that Π_A is a bijection from K_A to the union of $\tau^0_{A_1}$ and the union of all τ^2_B for which $B_1 = A$.

We allow a permutation π of the set of atoms to induce a permutation of the entire universe by the rule $\pi(A) = \pi``A$.

An A-allowable permutation is a permutation of atoms whose action fixes each τ^0_C for any C and each K_B for $B \ll A$. An $\emptyset$-allowable permutation is simply called an allowable permutation.

A small well-ordering of atoms and near-litters is called a support. An object X has A-support S iff S is a support and each A-allowable permutation π such that $\pi(S) = S$ also satisfies $\pi(X) = X$.

The collection τ^2_A consists exactly of those subsets of τ^1_A which have A-supports.

This actually completes the definition of the sets τ^2_A, mod the choice of the maps K_B for $B \ll A$, as long as we can verify that τ^2_A is of size μ in the ambient set theory.

The role of the maps χ_α is to provide an isomorphism between sets τ^2_A and $\tau^2_{A\cup B}$ when all elements of B dominate all elements of A: $\chi_{\texttt{max}(A)}$ witnesses an isomorphism between τ^2_A and $\tau^2_{A\setminus\{\texttt{max}(A)\}}$, and iteration of this fact gives the stated result.

The collection of all objects with $\emptyset$-supports (hereinafter supports) is a model of ZFA by the usual results about FM constructions.

τ^1_A is the power set of τ^0_α in the FM model defined by B-permutations for any B with $A \ll B$ (this is a statement which requires verification, but should not be difficult to believe).

τ^2_A is the power set of τ^1_A in the FM model determined by A-allowable permutations (this is evident from the way it is defined) and so is the double power set of τ^0_A. We make the claim to be verified that the subsets of τ^1_A with supports are the same as the subsets with A-supports, so in fact τ^2_A is the power set of τ^1_A in the FM model determined by all allowable permutations.

We verify that (subject to claims which need to be verified later) we can show that $\tau(A) = |\tau^2_A|$ defines a tangled web in the FM model determined by all allowable permutations.

Obviously $|K_A| \leq |\tau^2_A|$, since elements of K_A are elements of τ^2_A. An element of K_A, the local cardinal of a litter, has the singleton of that litter as a support. Further, in fact $2^{|K_A|} \leq \tau^2_A$, because subsets of K_A are in one to one correspondence with their set unions, which are elements of τ^2_A, because K_A is a pairwise disjoint collection. Because of the existence of the map Π_A, we have $|\tau^0_{A_1}| \leq |K_A|$ and $|\tau^2_B| \leq |K_A|$ when $B_1 = A$. We define $\exp(\kappa) = 2^\kappa$.

The inequalities above further give $\exp(|\tau^0_{A_1}|) \leq |\tau^2_A|$ and $\exp(|\tau^2_B|) \leq |\tau^2_A|$ when $B_1 = A$, so $\exp(|\tau^2_A|) \leq |\tau^2_{A_1}|$

Further, we get $\exp^2(|\tau^0_{A_1}|) = \tau(A_1) \leq \exp(|\tau^2_A|) = \exp(\tau(A))$.

and $\exp(|\tau^2_A|) = \exp(\tau(A)) \leq |\tau^2_{A_1}| = \tau(A_1)$ (where $|A| \geq 2$), so we have the naturality property of a tangled web for τ.

The natural model of TST_n with base type τ^2_A, where $n < |A|$, is sent by the composition of χ_α's determined by the elements of A_n to the natural model of TST_n with base type $\tau^2_{A\setminus A_n}$, and the χ_α's are external isomorphisms, so the first order theory of these models is the same. The reason for this is that the size of type $i < n$ in the first model is internally seen to be the same as that of $\tau^2_{A_i}$, and the second is internally seen to be the same size as that of $\tau^2_{(A\setminus A_n)_i} = \tau^2_{A_i\setminus A_n}$, and independently of the value of i the same composition of χ_α's serves as an external isomorphism. This verifies the elementarity property of τ.

This is an outline of how the proof works. What remains is the careful analysis of the way allowable permutations work which serves to verify that each set τ_A^2 is of size μ, that the power set of τ_A^0 in the FM models is τ_A^1, and that the subsets of τ_A^2 in the model determined by A-allowable permutations are the same as those in the model determined by all allowable permutations. What is required is results showing that allowable permutations act quite freely, and that is a further story.