Quine's NF—60 years on

Thomas Forster

July 18, 1998

Sixty years ago in this journal, the distinguished American philosopher W.V. Quine published a novel approach to set theory. The title was *New Foundations for Mathematical Logic* [6]. The diamond anniversary is being commemorated by a workshop in Cambridge (England) and comes at a time of rapid increase of interest in the alternatives to the hitherto customary *Zermelo-Fränkel* set theory, which promises a new lease of life for the axiomatic system now known as '*NF*'; its creator remains in good health too. Although he is best known to a wider public for his philosophical writings, his most enduring and most concrete legacy for the next fifty years may well turn out to be his most mathematical: he gave us *NF*.

Set theory is the study of sets, which are the simplest of all mathematical entities. Let us illustrate by constrasting sets with groups. Two distinct groups can have the same elements and yet be told apart by the way those elements are related. Sets are distinguished from all other mathematical fauna by the fact that a set is constituted solely by its members: two sets with the same members are the same set. To use a bit of jargon from another age, sets are properties *in extension*. As a result, all set theories have the axiom of extensionality: $(\forall xy)(x = y \longleftrightarrow (\forall z)(z \in x \longleftrightarrow z \in y))$: they differ in their views on which properties have extensions.

Since set theory first sprang on the scene about a hundred years ago there has been a tendency to attempt to use this simplicity to simplify and illuminate the rest of mathematics by translating (perhaps a better word is *implementing*) it into set theory. After all, if we can represent all of mathematics as facts about these delightfully simple things, some facts about mathematics might become clear that would otherwise remain obscure. This same simplicity means that set theory is always a good topic on which to try out any new mathematical idea.

Early twentieth century mathematicians used the expression "The Crisis in Foundations". This crisis had many causes and—despite the disappearance of the expression from contemporary speech—has never really been resolved. One of its many causes was the increasing formalisation of mathematics, which brought with it the realisation that the paradox of the liar could infect even mathematics itself. This appears most simply in the form of Russell's paradox, appropriately in the heart of set theory. At first blush one might think that where sets are concerned any intension has an extension: this is the axiom of naïve set existence. For any property of sets there is a set containing precisely the sets with that property, all of those and no others. This leads rapidly to Russell's paradox, the paradox of the class of all sets that are not members of themselves. This is the Russell class. Is it a member of itself? Well, if it is it isn't and if it isn't it is. This is Russell's paradox. The *aperçu* that leapt to mind was that the problem is something to do with the possibility of sets being members of themselves, or to do with defining sets in terms of membership in themselves. Although these two might sound like two formulations of the same insight, they nevertheless lead to radically different resolutions, and to two traditions in set theory represented by Zermelo-Fränkel set theory (often just called "set theory" by its votaries, and universally abbreviated to 'ZF') and Quine's *NF*, which is our primary concern here.

According to the first view, the source of the trouble manifested in Russell's paradox is thinking of sets as things that even *might* be members of themselves. This critique gives rise to a conception of set (usually called the *cumulative hierarchy* conception) that is very easy to explain to people in a modern computer science culture: it is simply the idea that sets form a recursive datatype:

> **The empty set is a set; any collection of sets forms a set; nothing else forms a set.**

This declaration carries with it a kind of induction principle, as recursive datatype declarations always do. If we have an assertion that is true of the empty set, and is true of any set x as long as it is true of all x's members, then it is true of all sets. This induction principle is $\in$-*induction* and is a theorem scheme of ZF. It has various consequences, of which one of the easiest to show is that no set is a member of itself. Clearly the empty set is not a member of itself. If no member of x is self-membered, then x cannot be self-membered either, otherwise x would be a self-membered member of x, contradicting the assumption that there aren't any. How does this way of conceiving sets help with Russell's paradox? Since no set is a member of itself, the collection of sets that aren't members of themselves would have to be the collection of all sets, and there can't be such a thing, since it would be a member of itself, and we've just used $\in$-induction to show that no set can be a member of itself.

If one had more space it would be natural to expand at this point on how the conception of sets as a recursive datatype gives rise to all (well, almost all!) the axioms of ZF by using $\in$-induction to show that the recursive datatype is closed under operations corresponding to those axioms. However, here the only reason for discussing ZF is to explain the difference between the conception of set that underlies it and the conception of set that underlies *NF*.

The *NF* conception of sets does not identify the problem behind Russell's paradox as a problem about the kind of set we are going to allow to exist, and therefore not as one that can be solved by banishing sets that do not belong to a nice recursive datatype. It locates the problem instead in *the way the sets are defined*. It does this by appeal to a concept of *type*, very closely related to the concept of type in modern typed programming languages such as ML. In an ML program, it must be possible to assign every variable a consistent type, subject to various typing rules; the same idea occurs in *NF*. Just as in ML, where one assigns types to variables in the context of a whole program, in *NF* one gives types to variables in a *formula*, and does not give a variable a type for life. In *NF* the types are natural numbers, and if the variable 'x' in a formula ϕ is given the type n and the subformula '$x \in y$' appears in ϕ, then we must give 'y' the type $n+1$. If '$x = y$' appears in ϕ then 'x' and 'y' must be given the same type. A formula is *stratified* if there is an assignment of types to variables that meets these constraints; otherwise it is *unstratified*. *NF*'s axioms are now very simply stated: (i) Extensionality; (ii) a scheme that says that the extension of a stratified formula is a set.

Let's try this on $\neg(x \in x)$. Clearly we will end up trying to give 'x' two distinct types and concluding that the formula is untyped. Therefore there is no axiom of *NF* saying that the collection of all sets that are not members of themselves is a set, and so, *prima facie*, no paradox. The other paradoxes are all held at bay in the same way. I am careful not to say that they are *avoided*, for it is an open question whether or not *NF* is consistent, but they are all held at bay in the sense that the obvious derivation for each paradox relies on a set-existence axiom that is not available in *NF* because the relevant formula is not stratified.

So far so good: stratification seems to prevent the usual paradoxes from being derivable, but are there any deep reasons why one would expect it to have this effect, or is it just a happy—and perhaps merely temporary—coincidence? Naturally people have tried to find reasons why stratification ought to work in this way, and it turns out that stratification is not a purely syntactical notion. To explain why, we need a device first used by Bernays and Rieger to prove the independence of the axiom of foundation from ZF. A model $\mathcal{M}$ of set theory is a class with a binary relation on it, typically written $\langle M, \in \rangle$. Now let π be a permutation of M, and associate with M a new relation, which holds between x and y precisely if $x \in \pi(y)$. If there is a universal set in the model $\langle M, \in \rangle$ then there is one in the new structure too, because if V was the universal set of $\langle M, \in \rangle$ then $\pi^{-1}(V)$ will be the universal set under the new dispensation. The assertion that there is a universal set is stratified, and it turns out that not only is the assertion that there is a universal set preserved by such redefinitions of the membership relation by permutations, but also *every* stratified assertion is thus preserved. (Subject to some small print the converse is true too: every sentence thus preserved is equivalent to a stratified formula.) Although this equivalence tells us that the apparently purely syntactical concept of stratification does have some semantical significance, it doesn't seem to tell us that this significance has anything to do with the avoidance of paradox. The clearest manifestation of this gap in our understanding is that our insight about the meaning of stratification has not yet given rise to a consistency proof for *NF*.

The feeling among modern *NF*istes is that this fact about stratified formulæ (which I like to think of as a *completeness theorem* since it identifies a semantical and a syntactic property) is nevertheless something that should be taken seriously. The argument runs like this: I said just now that a model of set theory is a set (M, say) with a binary relation (R, say) on it. For present purposes we want to think of a model of set theory as a set M of atoms (things with no internal structure) associated with an injective map $i : M \hookrightarrow \mathcal{P}(M)$, from M into the power set of M, so that the original R associated with M can be recovered as the relation $a \in i(b)$ (where '$\in$' is the membership relation of the real world in which we who are contemplating the model reside). We can think of i as a coding function: each $a \in A$ "codes" a subset of A, namely $\{x \in A : x \in i(a)\}$. We know from Cantor's theorem (every set is smaller than its power set) that not every subset of A can be coded by a member of A, so in constructing a model of set theory we have to leave some sets of atoms uncoded by atoms. A decision on what injection i to associate with A is (among other things) a decision about which collections of atoms are to be sets. Now revisit the idea of the "permutation models" of the preceding paragraph. If π is again a permutation of A then we can define a to be a member of b not if (as at the start of this paragraph) $a \in i(b)$ but instead if $a \in i(\pi(b))$, and we obtain another model of set theory. What is the difference between these two models? Well (since i and $i \circ \pi$ have the same range) they have made the same decision about which classes of atoms are to be sets, but different views on how that decision is to be implemented: the same collections of atoms are to be sets of the model, it is just that they are not necessarily going to be coded by the same elements of A as before. Accordingly the general feeling among *NF*istes is that stratification is the syntactical arm of a gang of concepts to do with what computer scientists call *implementation-invariance*.

But this is all very unhistorical. Let us go back to the years following 1937. *NF* was born in interesting times, and the West had other things on its mind during *NF*'s youth. The first really interesting development did not take place until 1953, when E.P. Specker in Zürich showed that *NF* refuted the axiom of choice and thereby proved the axiom of infinity [7]. This result was a most mysterious and disquieting one, best approached in the context of another result of Specker's, nine years later, that is in many ways more illuminating.

Specker's 1962 paper [9] connects *NF* with Russellian type theory in a way that neatly turns back the clock about 50 years. The syntax of Russell's type theory is very nasty, but the elements needed to tell its story can be recounted relatively easily. In Russell's type theory, as simplified by Ramsey, every set belongs to a *type*. The bottom type is a type of atoms, and thereafer type $n+1$ consists of sets of things of type n. Every variable of the theory is constrained to range over one level only. Accordingly no allegation that the collection of all sets that aren't members of themselves is a set can even be *formulated* in this sort of theory, let alone proved. That fact was the attraction; there are of course drawbacks as well. One is that we thereby chuck out the baby with the bathwater, in the sense that as well as rendering unsayable things like the existence of the Russell class we also make certain apparently entirely innocent things unsayable as well. A specific consequence is that the Russell-Ramsey theory makes all sorts of assertions that look very similar but are actually distinct, even though in some sense one feels that they ought not to be. For example (according to Russellian type theory) there is no single empty set but an empty set at each type. The language does not enable us to say anything like $(\exists x)(\forall y)(y \notin x)$. But it can say $(\exists x_1)(\forall y_0)(y_0 \notin x_1)$, $(\exists x_2)(\forall y_1)(y_1 \notin x_2)$, $(\exists x_3)(\forall y_2)(y_2 \notin x_3)$ ...and so on, where the subscripts are type subscripts. The language clearly has an endomorphism executed as follows: take a formula, increase all the type subscripts in it by 1. The result is a new formula, written 'ϕ^+' if the first formula was 'ϕ'. What is the relation between ϕ and ϕ^+? In [8] Specker drew a parallel with projective geometry, which also has an automorphism like this. By interchanging 'point' and 'line', and interchanging 'lie on' with 'meet at' one can transform an assertion ϕ of projective geometry into another assertion of projective geometry, which is standardly called the *dual* of the first, and is written $\hat{\phi}$. It is standard that the dual of an axiom of projective geometry is another axiom. By induction on proofs one shows that the dual of a theorem is a theorem. But is $\hat{\phi} \longleftrightarrow \phi$ a theorem? It is not obvious one way or the other. In the case of projective geometry the story has a neat solution and a happy ending (the scheme $\phi \longleftrightarrow \hat{\phi}$ is equivalent to Desargues' theorem), but in the type theory case it is more interesting, and not just because now the '+' operation is not an involution. It is certainly the case that ϕ^+ is an axiom whenever ϕ is, and ϕ^+ is a theorem whenever ϕ is, but is $\phi \longleftrightarrow \phi^+$ always a theorem? The example of the infinitely many statements saying that there is an empty set at each type is one that suggests very strongly that $\phi \longleftrightarrow \phi^+$ ought to be a theorem!

It turns out that the scheme $\phi \longleftrightarrow \phi^+$ is *not* a theorem of Russellian type theory but that it is consistent with Russellian type theory if and only if *NF* is consistent: this is Specker's 1962 theorem. This is very fitting when one reminds oneself of Quine's thinking behind the set existence axiom of NF. Quine's view—expressed in this MONTHLY 60 years ago—was that the type discipline that banished the paradoxes from type theory did so by making it impossible to formulate certain set existence axioms (like that giving the Russell class), and that making multiple copies—one at each type—of apparently perfectly nonproblematic sets like the empty set is an unwanted side effect and not part of the solution. If we can avoid some of this duplication by means of judicious polymorphism then this is all to the good. The result was that Quine kept the type distinctions but instead of enforcing them at the level of syntax (so that '$x \in x$' would be illformed, as in Russellian type theory) enforced them merely at the stage of axioms of set existence, so that '$x \in x$' is wellformed, but its extension is not a set. A modern way to describe this development is to say that Quine obtained *NF* from Russellian type theory by relaxing its syntactic constraints by a bit of polymorphism, and that Specker's 1962 theorem makes this fact formal and explicit.

One consequence of Specker's discovery was the involvement of proof theory in *NF* studies. Any proof in *NF* of a stratified formula corresponds to a proof of a version of that formula (with type subscripts glued on) in Russellian type theory with a scheme of polymorphism: "from $\vdash \phi$ deduce $\vdash \phi^+$ and *vice versa*". This interchangeability relates the proof theory of *NF* to the proof theory of type theory and thereby places *NF* studies firmly in the mainstream of modern theoretical computer science. Once *NF* has been placed in such a context, it is natural to think about what happens to the ideas that gave rise to its birth if they are approached constructively. It is then natural in turn to see if the strange derivation of the axiom of infinity works from a constructive standpoint. It turns out that there is a sensible constructive version of *NF* in which we can prove that it is not the case that every set is finite, but (since constructively $\neg\forall xp$ is not the same as $\exists x\neg p$) we cannot—apparently—prove that there is an infinite set. When working with classical logic we are of course not hampered in this way, and if we can show that not every set is finite then V, the universe, is certainly infinite. Now according to *NF* V is a set (it is the extension of the expression '$x = x$' which is certainly stratified) and so too is its quotient under the equivalence relation "is the same size as". This quotient will also be infinite, and it will give us an implementation of the natural numbers. The contrast between the classical case and the constructive case, where although we can prove that not every set is finite, there doesn't appear to be any one set whose infinitude can be proved (and so we apparently cannot obtain an implementation of the natural numbers), suggests that it may be possible to prove the consistency of constructive *NF* by much simpler methods than will be needed to prove the consistency of *NF* itself.

There are other subsystems of *NF* for which we can in fact do more than merely piously hope for consistency proofs. Most of these achieve their consistency by restricting the number of comprehension axioms in one way or another. For example NF_2 has axioms to say that the universe is a boolean algebra under $\subseteq$ and that $\{x\}$ is always a set; *NF*O has in addition an axiom saying that $\{y : x \in y\}$ is a set. (The operation sending x to $\{y : x \in y\}$ enables us to show by induction on ϕ that $\{x : \phi(x, y_1 \ldots y_n)\}$ is a set as long as ϕ is stratified and quantifier-free, and it is actually an $\in$-isomorphism!) NF_3 allows $\{x : \phi\}$ as long as the corresponding set existence axiom can be stratified with no more than 3 types. There is also a pair of theories arising from a *third* version of the circularity critique: perhaps it is necessary not only to create sets in order (as we do in the cumulative hierarchy conception) so that each set consists only of sets created earlier, but also to restrict the ways in which we specify sets so that we can form $\{x : \phi\}$ only if ϕ not only does not hold of things created later, but does not even *quantify over* sets created later. The idea is that we should be allowed to form $\{x : \phi\}$ only if checking that x has the property ϕ does not involve examining sets we have not yet created. Set existence axioms obeying such a constraint are said to be *predicative* and it has been known for a long time that adding predicativity constraints makes consistency much easier to prove.

But the most interesting subsystem of *NF* doesn't arise in this way and was totally unexpected. This was *NF*U, uncovered by R.B. Jensen in 1969. If one weakens the extensionality axiom that is so central to set theory to allow for distinct empty sets ('U' for "Urelemente" which is what set theorists call empty sets: they are certainly very hard to tell apart!) but retains it for nonempty sets one obtains the system *NF*U. The corresponding manœuvre in ZF results in a system which is equiconsistent with ZF and was—before the development of forcing by Cohen in the 60's—used for independence proofs for the axiom of choice and the like. When we weaken *NF* to allow urelemente the effect is dramatically different: *NF*U is provably consistent and is very weak indeed, too weak to prove the axiom of infinity.

One could view the consistency of *NF*U merely as a vindication of Quine's insight that the type disciplines are enough by themselves to banish the paradoxes, even if we flirt with danger by playing with a bit of polymorphism, as does Holmes [3]. Although it certainly *is* such a vindication, it raises bigger questions than it answers. After all, if type disciplines are enough to put paradox to flight even when relaxed with polymorphism, why is there this dramatic difference in strength between *NF* with and without atoms? Clearly there is something else going on. (There is even the ghastly and largely unspoken possibility that the consistency of *NF*U might have nothing to do with stratification at all, but is purely the result of weakening extensionality (and thereby betraying set theory) and that even though *NF*U is consistent, *NF* itself isn't.)

But even if we do not yet understand clearly why *NF*U is so much weaker than *NF*, we can at least start to put this new system to use [4]. There is for the moment a great interest in alternatives to ZF, driven by the feeling that certain structures with non-wellfounded relations on them ought to be represented by sets. (A relation R on a set X is wellfounded if and only if for every nonempty subset $X' \subseteq X$ $(\exists y \in X')(\forall x \in X')(\neg(R(x,y)).)$ For a long time the standard implementation of ordinal numbers in ZF has been one that arranges for the (wellfounded) relation $<$ between ordinal numbers to be implemented by $\in$, and the idea is abroad that *all* binary relations between mathematical objects of interest should be thus representable by $\in$ between the sets chosen to implement those mathematical objects. Under the recursive datatype conception of sets (as in ZF) we can prove easily that $\in$ is a wellfounded relation on the universe of all sets. Consequently there is no possibility of representing the kind of illfounded relations that appear in computer science as relations between sets of ZF.

What is a suitable framework for this? A fashionable candidate about which a lot has been written recently is ZF with "antifoundation" axioms, of which a racy and entertaining treatment can be found in the recently published book [1]. Antifoundation axioms ensure that all binary relations between mathematical objects of interest are representable by $\in$ between the sets chosen to implement those mathematical objects. In a way this is a very unidiomatic thing to do to ZF. As we noted earlier, the recursive datatype conception of sets entails that $\in$ is a wellfounded relation. It is surely perverse to develop an axiomatic set theory on the basis of one conception of set, and then throw away that conception by adopting axioms that are incompatible with it—thereby rendering suspect all the axioms it gave rise to. If we are to postulate sets that are forbidden by the recursive-datatype conception, then there is no point in looking to axioms arising from that conception to tell us how those sets are going to behave. Surely it makes more sense to have axioms of set existence that never owed anything to that conception in the first place. Such a set of axioms is to be found in *NF*U.

Can *NF*U in addition provide a set theoretic framework containing $\in$-copies of all the structures we can describe, as postulated by the antifoundation axioms? It turns out that for various technical reasons antifoundation axioms are not consistent with *NF*U as they stand. They need to be restricted to hereditarily *small* sets. (A set is hereditarily small if and only if it is a small set of hereditarily small sets.) What is a small set? Fortunately there is an *embarras de richesse* of direct concepts of smallness: we could say that x is small if and only if x is wellordered, or if x is the same size as a wellfounded set, or x cannot be mapped onto the universal set, or is smaller than its power set. These last two seem a bit odd, but are actually quite natural in the context of *NF*U. According to *NF*U the universe is a set. Therefore Cantor's theorem, which says that every set is smaller than its power set, must fail. But it succeeds for some sets, and these typically tend to be smaller than those for which it fails. A slightly smoother notion is *strongly cantorian*. A set x is strongly cantorian if and only if the restriction of the singleton function to x is a set. Theorems of Jensen [5] and Holmes [3] tell us that the hereditarily strongly cantorian sets can be almost any ZF-style model we want. A place to look for substructures of models of *NF*U in which every set is small and antifoundation axioms are true would perhaps be the *greatest* fixed point for the operation $x \mapsto$ the set of small subsets of x. The least fixed point consists entirely of wellfounded sets and satisfies foundation rather than antifoundation.

There is no space in a brief retrospective like this to give adequate pointers to all the relevant literature, and I am uncomfortably aware that the work of my *Doktorvater* Maurice Boffa, the unofficial head of the Belgian school of *NF*istes is underrepresented in this survey, as is his collaboration with Marcel Crabbé and his rôle in furthering *NF* studies by supervising André Pétry and Roland Hinnion. Nobody likes to appear to be promoting his own work unduly, but sadly it really is true that the only book-length treatment of *NF* is [2]. This book also contains treatments of permutation models and all the subsystems of *NF* mentioned in this article. Fortunately for readers who have access to the web there is also Randall Holmes' *NF* website at `http://math.idbsu.edu/faculty/holmes.html`, which contains an exhaustive bibliography, links to other workers on *NF* and Holmes' introduction to *NF*U.

References

1. Barwise, K.J.J. and Moss, L. *Vicious Circles.* Cambridge University Press, 1996.
2. Forster, T.E. *An Essay on Set Theory with a Universal set.* second edition, Oxford Logic Guides, Oxford University Press, 1995
3. Holmes, M.R. The set theoretical program of Quine succeeded (but nobody noticed). *Modern Logic* **4** (1994) 1–47.
4. Holmes, M.R. Elementary set theory with a universal set, Cahiers du centre de Logique **10** to appear.
5. Jensen, R.B. On the consistency of a slight (?) modification of Quine's NF. *Synthese* **19** (1969) 250–63.
6. Quine, W.v.O. New foundations for mathematical logic. *Amer. Math. Monthly* **44** 91937) 70–80.
7. Specker, E.P. The axiom of choice in Quine's new foundations for mathematical logic. *Proc. Nat. Acad. Sci. U.S.A.* **39** (1953) 972–5.
8. Specker, E.P. Dualität. *Dialectica* **12** (1958) 451–65.
9. Specker, E.P. Typical ambiguity. In *Logic, methodology and philosophy of science.* Ed E. Nagel, Stanford University Press, 1962.