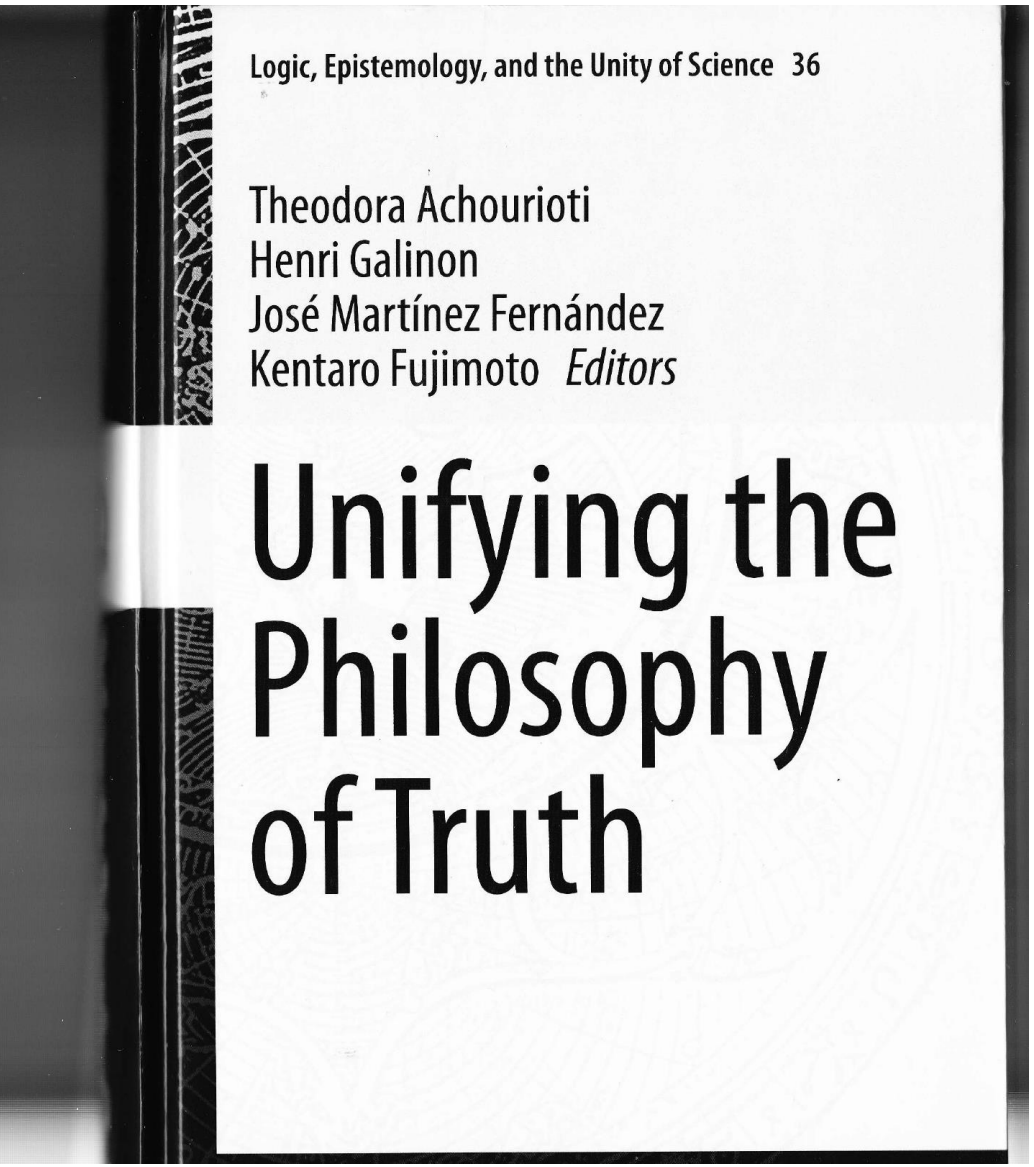
Logic, Epistemology, and the Unity of Science 36
Theodora Achourioti
Henri Galinon
José Martínez Fernández
Kentaro Fujimoto Editors
Unifying the Philosophy of Truth

Tarski, A. (1955). A lattice–theoretical fixpoint theorem and its applications. *Pacific journal of Mathematics* ,5(2), 285–309. http://projecteuclid.org/euclid.pjm/1103044538.
van Vugt, F. (2009). What makes a sentence be about the world? Master's thesis, Cogmaster, Ecole Normale Supérieure, Paris (France).
Yablo, S. (1993). Paradox without self-reference. *Analysis*, 53(4), 251–252.

Chapter 19
On Stratified Truth

Andrea Cantini

Abstract Is there a consistent axiomatization of a stratified form of the Tarskian hierarchy, where stratification is meant in the sense of Quine's *New Foundations* NF? In the following we propose a system of truth and abstraction, which might be regarded as an answer to the problem.

19.1 Stratified Truth: Introduction

Feferman recently[1] raised the question of finding a consistent axiomatization of the Tarskian hierarchy, where stratification is understood in Quine's sense. Some years ago in Cantini 2004 we sketched a reconstruction of a truth theory, based on the stratification discipline and aimed at a discussion of the Russellian paradox about propositions and sets.[2]

In the following we propose a modification of Cantini 2004, which might be regarded as partially approaching a solution to Feferman's problem. The basic idea is to devise a theory of truth SFT (= Stratified Fregean Truth) with strong expressive power, which is based upon *stratification as a means to achieve consistency*. As we shall see, stratification is exactly specified by the syntax of our system and validated by a set–theoretic interpretation in Quine's NF.

Let us anticipate a few informal considerations. First of all, the inspiring idea is to define a sort of analogue to the notion of Frege structure (in the sense of Aczel

This paper arises from the slides for the talk *Marginalia to self-referential truth*, presented at the Conference on *Axiomatic Theories of Truth* (New College, University of Oxford, September 19–20, 2011). We wish to thank the organizers for the nice hospitality and the stimulating environment. The research is supported by MIUR, under the project *Thinking and Computing, PRIN 2008* and within the frame of the University of Florence local research unit, sub-project *Abstraction and computation: logical and epistemological aspects.*

[1] In his talk at the Princeton Conference *Pillars of Truth*, April 8–10, 2011.

[2] See Russell's Principles of Mathematics, appendix B; the paradox is also known as the Russell–Myhill paradox.

A. Cantini
Dipartimento di Lettere e Filosofia, Sezione di Filosofia,
Università degli studi di Firenze via Bolognese, 52 50139 Firenze, Italy
e-mail: andrea.cantini@unifi.it

Aczel 1980), whose existence does not depend on lambda calculus, but on assuming a version of the discipline of types as codified by the practice of stratification. The universe is a kind of *abstract logical system* with a built-in *reflection mechanism*: if A represents a given proposition, there will be an object represented by $[A]$, which can be transformed into *a statement of a higher level* by applying the truth predicate T to $[A]$. But, in order to achieve consistency, we are forced to keep track of this natural level stratification: if $[A]$ is given type level i, $T([A])$ is assigned type $i+1$. We underline that truth is here regarded as a predicate T having a wider domain than usual truth predicates: T applies *not only to sentences of an inductively defined formal language*, but in general to objects of the given universe, which will possibly model – or play the role of – propositions (henceforth termed as *propositional objects*). Properly speaking, propositions remain defined *after* the semantical notion of truth is available (see definition 2.1), but we identify certain structures as possible propositional objects (see below definition 1.5). In this connection, the universe can be regarded as a sort of abstract syntax, where objects can be used to convey meanings. Thus we assume that the universe is closed under constructors, that are intended to represent *logical operations* and build *propositional objects*, and under an abstraction operator forming *predicative objects*. In particular, the universe includes objects encoding formulas of our truth language $\mathcal{L}_T$ and definable predicates thereof: if A is a sentence of $\mathcal{L}_T$, a map $A \mapsto [A]$ for forming propositional objects and an abstraction operation $x, A \mapsto [x|A]$ for building predicates are available, in such a way that the free variables of $[A]$ are the same as the free variables of A, and the free variables of $[x|A]$ are the free variables of A minus x (in short, $FV(A) = FV([A])$ and $FV([x|A]) = FV(A) - \{x\}$).

No type restriction is imposed in forming $[A]$ and $[x|A]$; but, as we shall see in the next subsection, the *use* of these expressions has to be suitably restricted according to the initial type-theoretic intuition.

If we exclude equality which is assumed as given, truth T is the only predicate of our language. But we also assume that there is a primitive operation $pred$, such that, roughly, $pred(t,s)$ expresses the fact that the object represented by s falls under the concept represented by t. T and $pred$ allow to express predication: if $[x|A]$ represents a predicate P defined by a given formula A, the result of the application of P to a, is rendered by $pred([x|A]), a)$. Hence the claim that a falls under P simply becomes the claim that $pred([x|A], a)$ *is true*, i.e. $T(pred([x|A], a))$.

One may wonder if self-referential constructions are, to a limited extent, allowed in the present framework. The answer is positive, but we stress that the present treatment is different from self-reference in the case of standard formal languages $\mathcal{L}$ (e.g. first order Peano arithmetic or ZFC), where one makes use of a substitution operation acting on Gödel numbers of $\mathcal{L}$-formulas and $\mathcal{L}$-terms. In the present framework, we assume that the universe is closed under a fixed point operator acting on terms, which depend *extensionally* on their parameters and are *homogeneously stratified*, i.e. arguments and values are assigned the same type. Roughly, the idea is that, if a propositional function F is extensional in a parameter x of given type i, then there

is a fixed point c of type i, i.e. such that $F(c) = c$. Of course, one has to clarify what is the meaning of *extensional in* x and the notion of type assignment in a formally untyped framework. Once clarified, these requirements are sufficient to sterilize self-reference and Liar's arguments: according to the basic intuition, if $tr(y)$ is the object representing the truth of y, $tr(y)$ is *type raising*, i.e. is assigned type one greater than y itself. Hence no fixed point of $tr(y)$ or of its negation will arise.

As to the justification, we shall see that self-reference follows by the set theoretic representation of logical constructors and as a consequence of (a variant of) the Knaster–Tarski theorem.

19.1.1 The Language and Its Stratification

Definition 1.1 The language $\mathcal{L}_T$ and its syntax. $\mathcal{L}_T$ includes:

- a unary predicate T for truth, a binary predicate $=$ for equality;
- binary function symbols id, $pred$, and; unary function symbols tr, neg, all;
- the binding operators $[-|-]$ (abstraction) and μ (fixed point).

id, $pred$, tr internally represent *basic constructors* (i.e. constructors for atomic formulas), while neg, and, all internally represent *logical constructors*.

Recall that, if E is an expression (i.e. term or formula) of the language, $FV(E)$ denotes the set of free variables of E.

Definition 1.2 We present a simultaneous inductive definition of the notions of (i) *term*, (ii) *term operative in a list* $\vec{x}$ *of parameters*, and (iii) *formula*:

- variables (and possibly individual constants when available) are terms;
- if x is a variable and $\vec{y}$ is a list of variables possible including x, then x is operative in $\vec{y}$;
- if t, s are terms, and f (g) is a binary (unary) function symbol, then $f(t,s)$ ($g(t)$) is a term;
- if t operative in $\vec{y}$ and s is operative in $\vec{z}$, then $and(t,s)$, $id(t,s)$ are operative in $\vec{y}, \vec{z}$, and $neg(t)$, $all(t)$ are operative in $\vec{y}$;
- if t is operative in $\vec{y}$, $pred(t,s)$ is operative in $\vec{y}$;
- if A is a formula, $[x|A]$ is a term such that $FV([x|A]) = FV(A) - \{x\}$;
- if $t(y,\vec{x})$ is operative in $y,\vec{x}$, then $\mu y t(y,\vec{x})$ is a term operative in $\vec{x}$, such that $FV(\mu y t(y,\vec{x})) = FV(t(y,\vec{x})) - \{y\}$;
- if t, s are terms, $T(t)$, $t = s$ are formulas; if A, B are formulas, then $\neg A$, $A \wedge B$, $\forall x A$ are formulas, and $FV(\forall x A]) = FV(A) - \{x\}$.

NB. The notion of *operative in* $\vec{x}$ corresponds to being extensional in $\vec{x}$, as informally hinted at in the introduction. As to specific examples, note that $neg(pred(y,x))$ is operative in y but not in x; $tr(y)$, $neg(tr(y))$ are not operative in y.

Of course, the language is suspiciously Fregean in the widest sense, and it is to be expected that we need some sort of restriction for governing truth and predication.

Indeed, in order to state the T-schema and the comprehension schema, we extend *the discipline of types* to arbitrary expressions E of the new language. The basic idea is that predication makes sense only in agreement with a suitable modification of Quine's stratification: informally, a predicate (represented by) t truly applies to s, where s is assigned type i, only if t is assigned type $i+1$. Similarly, truth has an *implicit hierarchical structure*: when we apply the predicate T to (a propositional object represented by) the term t, T must be assigned a level higher than the type assigned to t.

Definition 1.3 (Stratification of terms and formulas.) If E is an expression, E is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence and to each T-occurrence of E, so that:

1. all free occurrences of the same variable in any subexpression of E have the same type;
2. in each expression of the form $pred(t,s)$ the type of t is one greater than the type of its argument s; $pred(t,s)$ is assigned the type of t;
3. each expression of the form $tr(t)$ is assigned a type one greater than the type of t; in each expression of the form $T(t)$ T is assigned a type one greater than the type of t;
4. in each expression of the form $t = s$, $id(t,s)$ the type of t is the same type as s; $id(t,s)$ is assigned the same type of t (and hence of s);
5. each expression of the form $neg(t)$, $all(t)$ is assigned the same type of t;
6. each expression of the form $and(t,s)$ is assigned the same type as the type of t, s (that must have received the same type);
7. each term of the form $[x \mid C]$ is assigned a type one greater than the type assigned to x, and all the free occurrences of x in C receive the same type;
8. in each expression of the form $\forall x A$, if x is free in A, then the free occurrences of x in A and the occurrence of x in $\forall x$ receive the same type;
9. each term of the form $\mu y t(y, \vec{x})$ is assigned the same type as y and t, and all the free occurrences of $\vec{x}$ in t receive the same type.

NB. Within the same statement, different occurrence of T can be assigned different type labels and this makes sense of the idea of *typical ambiguity* in the semantical framework we are dealing with. Observe also that the definition of stratification imposes a *homogeneity condition* on $and(t,s)$; a semantical justification is to be found in the Quinean interpretation developed in Sect. 19.3 below.

Definition 1.4 We then inductively introduce $A \mapsto [A]$ with $FV(A) = FV([A])$:

- $[t = s] := id(t,s)$;
- $[T(t)] := tr(t)$
- $[\neg A] := neg([A])$;
- $[A \wedge B] := and([A],[B])$;
- $[\forall x A] := all([x \mid A])$

NB. It is not always true that, if A is stratified, then so is $[A]$; for instance, $T(x) \wedge T([Tx])$ is stratified (e.g. assign 0 to x, 1 to the first occurrence of T and 2 to the second); but $[T(x) \wedge T([Tx])]$ is not, as it fails to meet the homogeneity condition required by *and*. Thus it is the very mechanism of associating propositional objects that in this context can ruin stratification.

Definition 1.5 [P-Form] If an object x is in the range of the logical constructors, then it is called a *P-form*:

$$Pfor(x) \Leftrightarrow \exists y(x = tr(y)) \vee \exists z(x = neg(z) \vee x = all(z)) \vee \vee \exists u \exists v(x = id(u,v) \vee x = and(u,v))$$

Roughly, a P-form is an object which is (possibly) apt to represent a proposition. Observe that $Pfor(x)$ is stratified (assign 1 to x, u, v, z, 0 to y).

19.1.2 Axioms of SFT

SFT consists of the classical logical calculus (say, Hilbert-style) with equality and, in addition, the following axioms.

1. Compositional T-axioms:

$$T(id(x,y)) \leftrightarrow x = y;$$
$$T(neg(id(x,y)) \leftrightarrow \neg x = y;$$
$$T(tr(x)) \leftrightarrow T(x);$$
$$T(neg(tr(x)) \leftrightarrow \neg T(x);$$
$$T(neg(neg(x)) \leftrightarrow T(x);$$
$$T(and(x,y)) \leftrightarrow T(x) \wedge T(y);$$
$$T(neg(and(x,y)) \leftrightarrow T(neg(x)) \vee T(neg(y));$$
$$T(all(f)) \leftrightarrow \forall x T(pred(f,x));$$
$$T(neg(all(f)) \leftrightarrow \exists x T(neg(pred(f,x)))$$

2. T-consistency:

$$\neg(T(a) \wedge T(neg(a)))$$

3. T is well-defined on predication:

$$T(pred(f,x)) \vee T(neg(pred(f,x)))$$

4. Stratified β-conversion: if A is stratified,

$$T(pred([x|A],u)) \leftrightarrow T([A[x := u]])$$
$$T(neg(pred([x|A],u))) \leftrightarrow T([\neg A[x := u]])$$

Roughly, this schema states that, insofar as stratified conditions and truth contexts are involved, predicate abstraction and predicate application behave as inverse to each other.[3]

5. Self-reference: if t is operative in the list $y, \vec{x}$ and stratified,

$$\forall \vec{x}(t(\mu y t(y,\vec{x}),\vec{x}) = \mu y t(y,\vec{x}))$$

6. P-form:

$$T(x) \to Pfor(x)$$
$$\neg Pfor(x) \to T(neg(x))$$

The P-form axioms grant that true objects lie in the range of logical constructors; furthermore, any object inaccessible to logical constructors is classified as (representing) False.

7. μ-Extensionality: if two terms operative in $\vec{x}, y$ and stratified, are pointwise equal, then the respective fixed points coincide:

$$\forall \vec{x}\forall y(t(y,\vec{x}) = s(y,\vec{x})) \to \forall \vec{x}(\mu y t(y,\vec{x}) = \mu y s(y,\vec{x}))$$

8. Basic constructors and logical constructors are injective but not surjective, and their images are disjoint. In details, if f, g are distinct basic or logical constructors, f unary and g binary, then:

$$f(x) = f(y) \to x = y$$
$$g(x,y) = g(u,v) \to x = u \wedge y = v$$
$$\forall x\forall y\forall z(f(x) \neq g(y,z))$$
$$\exists x \neg Pfor(x)$$

Remark 1 The equivalence between $T(neg(tr(x)))$ and $\neg T(x)$ is *strongly non-kripkean* and makes the truth predicate closer to its classical counterpart. A similar comment holds for the clause involving predication.

19.1.3 Stratified Truth in SFT

The truth predicate is not only provably partial:

[3] Of course, one might simply postulate β-conversion at the object level, i.e. if A is stratified,

$$pred([x|A],u) = [A[x := u]]$$

Then the schemata would be trivially derivable. The reason is that we do not know how to prove its consistency.

Proposition 1.6 SFT *proves, for some closed term* L:

$$\neg T(L) \wedge \neg T(neg(L))$$

Moreover:

$$T(\neg T(L) \wedge \neg T(neg(L)))$$

Proof By self-reference choose $L = neg(L) = \mu y.neg(y)$. Then apply logic, T-consistency and the axioms relating T with tr, neg and and. □

Hence, not surprisingly, T is provably internally undefined on (the simplest variant of) the Liar; but, interestingly, T *internally believes* this fact.

Lemma 1.7 *(Compositional schemata) If A and B are arbitrary,*

(i) $T[\neg A] \leftrightarrow \neg T[A]$;
(ii) $T[A \wedge B] \leftrightarrow T[A] \wedge T[B]$;
(iii) $T[\forall x A] \leftrightarrow \forall x T[A(x)]$, *provided A is stratified.*

Proof As to (i), proceed by induction on A using the fact that T is well-defined on identities, predication and truth. (ii): it follows from the sixth axiom. (iii): apply stratified β-conversion. □

Proposition 1.8 *(Uniform stratified T-schema). If A is stratified,* SFT *proves:*

$$\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x})) \tag{19.1}$$
$$T[\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x}))] \tag{19.2}$$

Proof We check by simultaneous induction on A

$$(T([A]) \leftrightarrow A) \wedge (T([\neg A]) \leftrightarrow \neg A)$$

If A is of the form $t = s$, $T(t)$, apply the corresponding axioms of SFT.

If A is of the form $B \wedge C$, even if $A \wedge B$ is stratified, $[A \wedge B]$ may be not. However, by $\forall$-instantiation of the compositional axiom about T and $\wedge$, we obtain

$$T([A \wedge B]) \leftrightarrow T([A]) \wedge T([B])$$

where the left hand side is not stratified in general, while the right hand side is stratified. Then we apply IH.[4]

If A is of the form $\neg B$, apply the compositional SFT-axioms involving negated $\wedge$, double negation, and IH.

Let us consider the case of a negated universal quantifier. Then we use the axioms relating T, $\neg\forall$, together with β-conversion and IH in the final step:

$$T([\neg\forall x A]) \leftrightarrow T(neg(all([x|A])))$$

[4] Henceforth IH stands for induction hypothesis in short.

Roughly, this schema states that, insofar as stratified conditions and truth contexts are involved, predicate abstraction and predicate application behave as inverse to each other.[3]

5. Self-reference: if t is operative in the list $y, \vec{x}$ and stratified,

$$\forall \vec{x}(t(\mu y t(y,\vec{x}),\vec{x}) = \mu y t(y,\vec{x}))$$

6. P-form:

$$T(x) \to Pfor(x)$$
$$\neg Pfor(x) \to T(neg(x))$$

The P-form axioms grant that true objects lie in the range of logical constructors; furthermore, any object inaccessible to logical constructors is classified as (representing) False.

7. μ-Extensionality: if two terms operative in $\vec{x}, y$ and stratified, are pointwise equal, then the respective fixed points coincide:

$$\forall \vec{x}\forall y(t(y,\vec{x}) = s(y,\vec{x})) \to \forall \vec{x}(\mu y t(y,\vec{x}) = \mu y s(y,\vec{x}))$$

8. Basic constructors and logical constructors are injective but not surjective, and their images are disjoint. In details, if f, g are distinct basic or logical constructors, f unary and g binary, then:

$$f(x) = f(y) \to x = y$$
$$g(x,y) = g(u,v) \to x = u \wedge y = v$$
$$\forall x\forall y\forall z(f(x) \neq g(y,z))$$
$$\exists x \neg Pfor(x)$$

Remark 1 The equivalence between $T(neg(tr(x)))$ and $\neg T(x)$ is *strongly non-kripkean* and makes the truth predicate closer to its classical counterpart. A similar comment holds for the clause involving predication.

19.1.3 Stratified Truth in SFT

The truth predicate is not only provably partial:

[3] Of course, one might simply postulate β-conversion at the object level, i.e. if A is stratified,

$$pred([x|A],u) = [A[x := u]]$$

Then the schemata would be trivially derivable. The reason is that we do not know how to prove its consistency.

Proposition 1.6 SFT *proves, for some closed term* L:

$$\neg T(L) \wedge \neg T(neg(L))$$

Moreover:

$$T(\neg T(L) \wedge \neg T(neg(L)))$$

Proof By self-reference choose $L = neg(L) = \mu y.neg(y)$. Then apply logic, T-consistency and the axioms relating T with tr, neg and and. □

Hence, not surprisingly, T is provably internally undefined on (the simplest variant of) the Liar; but, interestingly, T *internally believes* this fact.

Lemma 1.7 *(Compositional schemata) If A and B are arbitrary,*

(i) $T[\neg A] \leftrightarrow \neg T[A]$;
(ii) $T[A \wedge B] \leftrightarrow T[A] \wedge T[B]$;
(iii) $T[\forall x A] \leftrightarrow \forall x T[A(x)]$, *provided A is stratified.*

Proof As to (i), proceed by induction on A using the fact that T is well-defined on identities, predication and truth. (ii): it follows from the sixth axiom. (iii): apply stratified β-conversion. □

Proposition 1.8 *(Uniform stratified T-schema). If A is stratified,* SFT *proves:*

$$\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x})) \tag{19.1}$$
$$T[\forall x(T([A(\vec{x})]) \leftrightarrow A(\vec{x}))] \tag{19.2}$$

Proof We check by simultaneous induction on A

$$(T([A]) \leftrightarrow A) \wedge (T([\neg A]) \leftrightarrow \neg A)$$

If A is of the form $t = s$, $T(t)$, apply the corresponding axioms of SFT.

If A is of the form $B \wedge C$, even if $A \wedge B$ is stratified, $[A \wedge B]$ may be not. However, by $\forall$-instantiation of the compositional axiom about T and $\wedge$, we obtain

$$T([A \wedge B]) \leftrightarrow T([A]) \wedge T([B])$$

where the left hand side is not stratified in general, while the right hand side is stratified. Then we apply IH.[4]

If A is of the form $\neg B$, apply the compositional SFT-axioms involving negated $\wedge$, double negation, and IH.

Let us consider the case of a negated universal quantifier. Then we use the axioms relating T, $\neg\forall$, together with β-conversion and IH in the final step:

$$T([\neg\forall x A]) \leftrightarrow T(neg(all([x|A])))$$

[4] Henceforth IH stands for induction hypothesis in short.

$$\leftrightarrow \exists u T(neg(pred([x|A], u)))$$
$$\leftrightarrow \exists u T([\neg A[x := u]])$$
$$\leftrightarrow \exists u(\neg A[x := u]) \equiv \neg \forall x A$$

The case of positive $\forall$ is similar. □

The stratified T-schema implies that T strongly deviates from the behaviour of self-referential truth predicates à la Kripke–Feferman, which cannot in general be applied to the truth axioms themselves, nor to *arbitrary* logical axioms. On the contrary, T *provably believes that it is two-valued and consistent*; further, it recognizes that each closure condition is also internally true.

Corollary 1.9

(i) SFT *proves:*

$$T([T(a) \vee \neg T(a)]);$$
$$T([\neg(T(a) \wedge T(neg(a)))])$$

(ii) Moreover, if Axiom *is an instance of a compositional T-axiom or T-welldefinedness,* SFT *proves* $T([Axiom])$.

Proof Observe that the consistency statement as well as *tertium non datur* for T and the compositional axioms are stratified; hence the claim is a consequence of the stratified truth schema. □

Remark 2 One may wonder whether the fixed point property can be extended, e.g. up to include the constructors tr, $pred$ and combinations thereof. It is immediate to see that the answer is negative. Indeed, assume that there exists e such that

$$e = neg(tr(e))$$

Then $T(e) \leftrightarrow T(neg(tr(e))) \leftrightarrow \neg T(e)$: contradiction! The reason is that, roughly, as we shall see in the model construction, neither $x \mapsto tr(x)$ nor $x \mapsto pred(y, x)$ are monotone (in the sense of set theoretic inclusion) with respect to x.

We conclude by showing that SFT proves that its truth predicate is indeed the fixed point of a natural positive operator. Let $\mathcal{V}(x, T)$ be the formula:

$$\begin{aligned}
&\exists v(\neg Pfor(v) \wedge x = neg(v)) \vee \\
\vee\ &\exists w_1((x = [T(w_1)] \wedge T(w_1)) \vee \\
&\vee (x = [\neg T(w_1)] \wedge \neg T(w_1))) \vee \\
\vee\ &\exists w_3(x = neg(neg(w_3)) \wedge T(w_3)) \vee \\
\vee\ &\exists w_4 \exists w_5((x = id(w_4, w_5) \wedge w_4 = w_5) \vee \\
&\vee (x = neg(id(w_4, w_5)) \wedge w_4 \neq w_5)) \vee \\
\vee\ &\exists w_6 \exists w_7(((x = and(w_6, w_7)) \wedge T(w_6) \wedge T(w_7)) \vee
\end{aligned}$$

$$\begin{aligned}
&\vee (x = neg(and(w_6, w_7)) \wedge (T(neg(w_6)) \vee T(neg(w_7))))) \vee \\
\vee\ &\exists f((x = all(f) \wedge \forall z T(pred(f, z))) \vee \\
&\vee (x = neg(all(f)) \wedge \exists z T(neg(pred(f, z)))))
\end{aligned}$$

Theorem 1.10 *(Fixed point principle)*

$$\forall x(T(x) \leftrightarrow \mathcal{V}(x, T))$$

Proof $\Rightarrow$: let $T(x)$. Then $Pfor(x)$. If $x = neg(y)$ and not $Pfor(y)$, clearly $\mathcal{V}(x, T)$. If $x = neg(y)$ but $Pfor(y)$, we distinguish several cases and we apply the T-compositional axioms from left to right. E.g. if $x = neg(id(u, v))$, then $\neg u = v$ and we conclude $\mathcal{V}(x, T)$.

$\Leftarrow$: if $\mathcal{V}(x, T)$, we again argue by cases using T-axioms from right to left. □

Corollary 1.11 *(Internal fixed point principle)*

$$T[\forall x(T(x) \leftrightarrow \mathcal{V}(x, T))]$$

Proof The statement of the fixed point theorem is stratified: hence apply the theorem and the stratified T-schema. □

19.2 On a Paradox About Propositions and Truth

The problem at issue is whether the predicate of *being a proposition* actually defines a genuine class (type) in Russell's sense. Assume that the answer is positive. Then, according to Russell 1903, a contradiction with a version of Cantor's theorem could be derived: there would exist an injection (see below definition 2.3 and lemma 2.4) from the collection of classes, whose elements are propositions, into the collection of propositions. It turns out that in the present situation stratification blocks the diagonalization leading to contradiction.

Definition 2.1

(i) $P(a) :\Leftrightarrow T(a) \vee T(neg(a))$;
(ii) $V := [x|x = x]$;
(iii) $imp(a, b) := neg(and(a, neg(b)))$;
(iv) $or(a, b) := neg(and(neg(a), neg(b)))$.

Pa formally represents the predicate "a is a proposition". We also define $a \subseteq b$ for $\forall u(T(pred(a, x)) \rightarrow T(pred(b, x)))$.

Proposition 2.2 *(SFT). The collection of all propositions is a proper subset of the universe:*

$$[x \mid P(x)] \subset V$$

Moreover P has the following closure properties:

$$P(id(y,x)) \wedge P(pred(y,x)) \wedge P(tr(x));$$
$$P(a) \wedge (T(a) \rightarrow P(b)) \rightarrow P(imp(a,b));$$
$$P(a) \wedge P(b) \rightarrow P(and(a,b)) \wedge P(or(a,b));$$
$$P(a) \rightarrow T([Pa]);$$
$$P(all(f)).$$

The first claim is a consequence of proposition 1.6. As to the remaining properties, apply the T-compositional axioms.

Note also that $P(all(f))$ implies $\forall x P(pred(f,x))$, i.e. *every set defines a propositional function.*

We now conclude by representing Russell's contradiction of appendix B (see Russell 1903) within the theory of propositions and truth.

Definition 2.3

$$\tau(f) := [P(all(f))]$$

By definition of the map $A \mapsto [A]$, the axiom that logical operators are injective, and proposition 2.2, we obtain:

Lemma 2.4 SFT proves:

$$P(\tau(f)) \wedge T(\tau(f));$$
$$\tau(f) = \tau(g) \rightarrow f = g$$

Informally, the operation τ is a well-defined injective map from sets into truths (and propositions). Also, observe that τ is a type-raising operation: if f is assigned type 0, $\tau(f)$ must receive type 1 (since it contains the operator tr). As a consequence, the formula

$$(\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f))$$

cannot be stratified.

Proposition 2.5 SFT *proves:*

$$\neg \exists d \forall x (T(pred(d,x)) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f))$$

Proof Assume by contradiction that there exists d such that

$$\forall x (T(pred(d,x)) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f,x)) \wedge x = \tau(f)))$$

If we choose $x := \tau(d)$, we get

$$T(pred(d,\tau(d))) \leftrightarrow (\exists f \subseteq P)(\neg T(pred(f,\tau(d))) \wedge \tau(d) = \tau(f)))$$

Assume $T(pred(d,\tau(d)))$. Then, as τ is injective, $d = f$ and hence $\neg T(pred(d,\tau(d)))$. Hence by minimal logic $\neg T(pred(d,\tau(d)))$. Also $d \subseteq P$: indeed, if x is arbitrary and $T(pred(d,x))$, then $x = \tau(f)$ for some f, whence by lemma 2.4, $P(\tau(f))$, i.e. $P(x)$. Hence by assumption on d, the contradiction $T(pred(d,\tau(d)))$. □

This is the solution of the paradox in appendix B: the diagnosis is that the paradoxical set d does not exist according to the discipline of types even in a liberalized Quinean sense.[5]

19.3 Embedding Stratified Truth in NF

Let $\mathcal{L}_s$ be the elementary set theoretic language, which comprises the binary predicate symbol $\in$. $\mathcal{L}_s$–terms are simply individual variables $(x, y, z, \ldots)$; prime formulas (atoms) have the form $t \in s$, $t = s$ (t, s terms). $\mathcal{L}_s$–formulas are inductively generated from prime formulas by means of sentential connectives and quantifiers. The elementary set theoretic language $\mathcal{L}_s^+$ is obtained by adding to $\mathcal{L}_s$ the abstraction operator $\{- \mid -\}$; $\mathcal{L}_s^+$–terms and formulas are then simultaneously generated. The clause for introducing class terms has the form: if φ is a formula, then $\{x \mid \varphi\}$ is a term where $FV(\{x \mid \varphi\}) = FV(\varphi) - \{x\}$ ($FV(E)$ is the set of free variables occurring in the expression E). Two terms (formulas) are called α–congruent if they only differ by renaming of bound variables; we identify α–congruent terms (formulas).

19.3.1 Stratified Comprehension

As usual for Quine's systems, we need *stratification*; we also define a restricted notion thereof, which is motivated by the consideration of "loosely predicative" class existence axioms.

---

[5] We underline that our formalization does not literally represent the paradox of the final section 500 of 2.2. For the reader's sake here is Russell's text:

> If m be a class of propositions, the proposition *"every m is true"* may or not be itself an m. But there is a one-one-relation of this proposition to m: if n be different from m, *"every n is true"* is not the same proposition as as *"every m is true"*. Consider now the whole class of propositions of the form *"every m is true"*, and having the property of not being members of their respective m's. Let this class be w, and let p be the proposition *"every w is true"*. If p is a w, it must possess the defining property of w; but this property demands that p should not be a w. On the other hand, if p is not a w, then p does possess the defining property of w, and therefore is a w. Thus the contradiction appears unavoidable.

(i) φ is *stratified* iff it is possible to assign a natural number (type in short) to each term occurrence [6] of φ in such a way that
- if $t \in s$ is a subformula of φ, the type of s is one greater than the type of t; if $t = s$ is a subformula of φ, the type of s is the same as the type of t;
- all free occurrences of the same variable in any subformula of φ have the same type;
- if x is free in ψ and $\forall x \psi$ is a subformula of φ, then the 'x' in $\forall x$ and the free occurrences of x in ψ receive the same type;
- if $t := \{x \mid \beta\}$ occurs in φ, x is free in β, then t is assigned a type one greater than the type assigned to x, and all the free occurrences of x in β receive the same type.

(ii) $\{x \mid \varphi\}$ is stratified if φ is stratified;

(iii) a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *loosely predicative* iff for some type $i \in \omega$, $\{x \mid \varphi(x, \vec{y})\}$ has type $i + 1$, no (free or bound) variable of $\varphi(x, \vec{y})$ is assigned type greater than $i+1$; a stratified term $\{x \mid \varphi(x, \vec{y})\}$ is *predicative* iff $\{x \mid \varphi(x, \vec{y})\}$ is loosely predicative and in addition no quantified variable of $\varphi(x, \vec{y})$ is assigned the same type as $\{x \mid \varphi(x, \vec{y})\}$ itself.

(iv) φ is $n + 1$–*stratified* iff φ is stratified by means of $0, \ldots, n$.

For instance, $\bigcup a = \{x \mid (\exists y \in a)(x \in y)\}$ is not loosely predicative, since it requires type 2, but $\bigcup a$ itself has type 1; $a \cap b = \{x \mid x \in a \wedge x \in b\}$ is predicative.

Definition 3.1 The system NF comprises:

(i) predicate logic for the extended language[7];

(ii) class extensionality: $\forall x \forall y(x =_e y \rightarrow x = y)$, where

$$t =_e s :\Leftrightarrow \forall x(x \in t \leftrightarrow x \in s)$$

(iii) stratified explicit comprehension SCA: if φ is stratified, then

$$\forall u(u \in \{x \mid \varphi(x, \vec{y})\} \leftrightarrow \varphi(u, \vec{y}))$$

[6] Individual constants included; these can be given any type compatible with the clauses below.

[7] If the abstraction operator is assumed as primitive, the extended logic contains the schema

$$\forall u(\varphi(u) \leftrightarrow \psi(u)) \rightarrow \{x \mid \varphi(x)\} = \{x \mid \psi(x)\}$$

Other systems

(a) NFP (NFI) is the subsystem of NF, where SCA is restricted to (loosely) predicative abstracts.

(b) NF_k (NFI_k, NFP_k) is the subsystem of NF (NFI, NFP), where (at most) k types are allowed for stratification.

Remark 3 By a theorem of Crabbè 1982, NFI is provably consistent in third order arithmetic. The details of the (different) consistency proofs for NFI can be found in Crabbè 1982 and Holmes 1995.

In order to carry out a Kripke-like construction in the NF-systems and to represent the syntax, we shall essentially exploit Quine's homogeneous pairing operation, which *does require extensionality* and the existence of a copy of the natural numbers. But it is not difficult to check that Quine's pairing is indeed well-defined already in NFI. First of all, the collection of Fregean natural numbers is a set in NFI. Define:

$$\emptyset = \{x \mid x \neq x\}$$
$$V = \{x \mid x = x\}$$
$$0 = \{\emptyset\}$$
$$a + 1 = \{x \cup \{y\} \mid x \in a \wedge y \notin x\}$$
$$Cl_{\mathcal{N}}(y) \Leftrightarrow 0 \in y \wedge \forall x(x \in y \rightarrow (x + 1) \in y)$$
$$\mathcal{N} = \{x \mid \forall y(Cl_{\mathcal{N}}(y) \rightarrow x \in y)\}$$

NFI proves the existence of $\mathcal{N}$; in fact, by inspection, all the above sets above are loosely predicative. Furthermore, we have, provably in NFI:

Lemma 3.2 *(NFI)*

$$Cl_{\mathcal{N}}(\{x \mid \varphi(x)\}) \rightarrow \mathcal{N} \subseteq \{x \mid \varphi(x)\} \quad (19.3)$$
$$(\forall x)(x \in \mathcal{N} \leftrightarrow x = 0 \vee (\exists y \in \mathcal{N})(x = y + 1)) \quad (19.4)$$
$$\emptyset \notin \mathcal{N} \wedge (\forall x \in \mathcal{N})(V \notin x) \quad (19.5)$$
$$(\forall x \in \mathcal{N})(x + 1 \neq 0) \quad (19.6)$$
$$(\forall x \in \mathcal{N})(\forall y \in \mathcal{N})(x + 1 = y + 1 \rightarrow x = y) \quad (19.7)$$

(In (19.3) $\{x \mid \varphi(x)\}$ must be loosely predicative).

Clearly $\mathcal{N}$ is infinite by (19.5) above. As to the proof, (19.5) holds in NFI + Union, as NFI+Union $\equiv$ NF, and NF proves (19.5) according to a famous result of Specker (1953). On the other hand, NFI + ¬Union implies (19.5) by Crabbè 1982. The claims (19.4), (19.3) with the Peano axioms are provable in NFI ((19.7) requires the second part of (19.5)).

Definition 3.3 (Homogeneous pairing; Rosser 1953)

$$\phi(a) = \{y \mid y \in a \wedge y \notin \mathcal{N}\} \cup \{y + 1 \mid y \in a \wedge y \in \mathcal{N}\};$$

$$\theta_1(a) = \{\phi(x) \mid x \in a\};$$
$$\theta_2(a) = \{\phi(x) \cup \{0\} \mid x \in a\};$$
$$(a,b) = \theta_1(a) \cup \theta_2(b);$$
$$Q_1(a) = \{z \mid \phi(z) \in a\};$$
$$Q_2(a) = \{z \mid \phi(z) \cup \{0\} \in a\}$$

The definitions above are (at most) loosely predicative and hence the universe of sets is closed under the corresponding operations, provably in NFI.

We below exploit the fact that Quine's pairing operation is $\subseteq$-monotone in both arguments: indeed, the definition of (a,b) is positive in a, b[8].

Lemma 3.4 *We have, provably in NFI:*

(i) $\phi(a) = \phi(b) \to a = b$;
(ii) $0 \notin \phi(a)$;
(iii) $\theta_i(a) = \theta_i(b) \to a = b$, *where* $i = 1, 2$;
(iv) $(x,y) = (u,v) \to x = u \wedge y = v$.
(v) the map $x, y \longmapsto (x,y)$ *is surjective and* $\subseteq$*–monotone in each variable, i.e.*

$$x \subseteq u \wedge y \subseteq w \to (x,y) \subseteq (u,w) \tag{19.8}$$

The proof hinges upon the properties of $\mathcal{N}$ and the successor operation (Rosser 1953).

Lemma 3.5 *(Fixed point) Let* $A(x,a)$ *be a formula which is positive in* a. *Assume that*

$$\Gamma_A(a) = \{x \mid A(x,a)\}$$

is loosely predicative, where x, a *are given types* i, $i+1$ *respectively. Then* NFI *proves the existence of a set* c *of type* $i+1$, *such that*:

- $\Gamma_A(c) \subseteq c$;
- $\Gamma_A(a) \subseteq a \Rightarrow c \subseteq a$.

The proof is standard: observe that the set

$$c := \{x \mid \forall d(\Gamma_A(d) \subseteq d \to x \in d)\}$$

is loosely predicative.

[8] We recall that a formula $A(x,a)$ is positive in a if every free occurrence of a in the negation normal form of A is located in atoms of the form $t \in a$, which are prefixed by an even number of negations and where $a \notin FV(t)$.

19.3.2 Generating Truth

We use Quine's pairing for representing logical constructors, and the fixed point lemma 3.5 to interpret the truth predicate.

Definition 3.6

$$\dot{\neg}x := (0,x);$$
$$x \dot{\wedge} y := (1,(x,y));$$
$$\dot{\forall} f := (2,f);$$
$$\dot{\in} xy := (3,(x,y));$$
$$\dot{=} xy := (4,(x,y))$$

Of course, the number labels above are natural numbers in the sense of lemma 3.2. We also write $[x = y]$ for $(\dot{=} xy)$. If $\{x\}$ denotes the singleton, we let

$$[x \in y] := \dot{\in}\{x\}y = y \cdot x$$

Under the dot-application, the universe of sets becomes an applicative structure. $y \cdot x$ is stratified only if y and x are given the types $i+1$ and i (respectively), and the result of applying y to x is one greater than the type of x.

We now model the Kripke–Feferman notion of self-referential truth within the abstract framework of Quine's set theory. First of all, in analogy with the notion of P-form, define

$$Pfr(x) \Leftrightarrow \exists u \exists v(x = [u \in v]) \vee \exists z(x = \dot{\neg} z \vee x = \dot{\forall} z) \vee$$
$$\vee \exists w_1 \exists w_2((x = [w_1 = w_2]) \vee (x = w_1 \dot{\wedge} w_2))$$

$Pfr(x)$ is stratified (assign 1 to x, v, z, w_1, w_2 and 0 to u). The truth predicate W is introduced as the fixed point of a stratified positive (in a) operator $\mathcal{T}(x,a)$, which encodes the recursive clauses for partial self-referential truth and is given by the formula

$$\exists y\,(x = \dot{\neg} y \wedge \neg Pfr(y)) \vee$$
$$\exists u \exists v \exists w\,[(x = [u \in v] \wedge u \in v) \vee$$
$$\vee (x = \dot{\neg}[u \in v] \wedge \neg u \in v) \vee$$
$$\vee (x = [v = w] \wedge\ v = w) \vee$$
$$\vee (x = [\neg v = w] \wedge\ \neg v = w) \vee$$
$$\vee (x = \dot{\neg}\dot{\neg} v \wedge\ v \in a) \vee$$
$$\vee (x = v \dot{\wedge} w \wedge\ v \in a \wedge w \in a) \vee$$
$$\vee (x = \dot{\neg}(v \dot{\wedge} w) \wedge\ (\dot{\neg} v \in a \vee \dot{\neg} w \in a)) \vee$$

$$\theta_1(a) = \{\phi(x) \mid x \in a\};$$
$$\theta_2(a) = \{\phi(x) \cup \{0\} \mid x \in a\};$$
$$(a,b) = \theta_1(a) \cup \theta_2(b);$$
$$Q_1(a) = \{z \mid \phi(z) \in a\};$$
$$Q_2(a) = \{z \mid \phi(z) \cup \{0\} \in a\}$$

The definitions above are (at most) loosely predicative and hence the universe of sets is closed under the corresponding operations, provably in NFI.

We below exploit the fact that Quine's pairing operation is $\subseteq$-monotone in both arguments: indeed, the definition of (a,b) is positive in a, b[8].

Lemma 3.4 *We have, provably in NFI:*

(i) $\phi(a) = \phi(b) \to a = b$;
(ii) $0 \notin \phi(a)$;
(iii) $\theta_i(a) = \theta_i(b) \to a = b$, *where* $i = 1, 2$;
(iv) $(x,y) = (u,v) \to x = u \wedge y = v$.
(v) the map $x, y \longmapsto (x,y)$ *is surjective and* $\subseteq$*–monotone in each variable, i.e.*

$$x \subseteq u \wedge y \subseteq w \to (x,y) \subseteq (u,w) \tag{19.8}$$

The proof hinges upon the properties of $\mathcal{N}$ and the successor operation (Rosser 1953).

Lemma 3.5 *(Fixed point) Let* $A(x,a)$ *be a formula which is positive in* a. *Assume that*

$$\Gamma_A(a) = \{x \mid A(x,a)\}$$

is loosely predicative, where x, a *are given types* i, $i+1$ *respectively. Then* NFI *proves the existence of a set* c *of type* $i+1$, *such that*:

- $\Gamma_A(c) \subseteq c$;
- $\Gamma_A(a) \subseteq a \Rightarrow c \subseteq a$.

The proof is standard: observe that the set

$$c := \{x \mid \forall d(\Gamma_A(d) \subseteq d \to x \in d)\}$$

is loosely predicative.

[8] We recall that a formula $A(x,a)$ is positive in a if every free occurrence of a in the negation normal form of A is located in atoms of the form $t \in a$, which are prefixed by an even number of negations and where $a \notin FV(t)$.

19.3.2 Generating Truth

We use Quine's pairing for representing logical constructors, and the fixed point lemma 3.5 to interpret the truth predicate.

Definition 3.6

$$\dot{\neg}x := (0,x);$$
$$x \dot{\wedge} y := (1,(x,y));$$
$$\dot{\forall} f := (2,f);$$
$$\dot{\in} xy := (3,(x,y));$$
$$\dot{=} xy := (4,(x,y))$$

Of course, the number labels above are natural numbers in the sense of lemma 3.2. We also write $[x = y]$ for $(\dot{=} xy)$. If $\{x\}$ denotes the singleton, we let

$$[x \in y] := \dot{\in}\{x\}y = y \cdot x$$

Under the dot-application, the universe of sets becomes an applicative structure. $y \cdot x$ is stratified only if y and x are given the types $i+1$ and i (respectively), and the result of applying y to x is one greater than the type of x.

We now model the Kripke–Feferman notion of self-referential truth within the abstract framework of Quine's set theory. First of all, in analogy with the notion of P-form, define

$$Pfr(x) \Leftrightarrow \exists u \exists v(x = [u \in v]) \vee \exists z(x = \dot{\neg} z \vee x = \dot{\forall} z) \vee$$
$$\vee \exists w_1 \exists w_2((x = [w_1 = w_2]) \vee (x = w_1 \dot{\wedge} w_2))$$

$Pfr(x)$ is stratified (assign 1 to x, v, z, w_1, w_2 and 0 to u). The truth predicate W is introduced as the fixed point of a stratified positive (in a) operator $\mathcal{T}(x,a)$, which encodes the recursive clauses for partial self-referential truth and is given by the formula

$$\exists y\,(x = \dot{\neg} y \wedge \neg Pfr(y)) \vee$$
$$\exists u \exists v \exists w\,[(x = [u \in v] \wedge u \in v) \vee$$
$$\vee (x = \dot{\neg}[u \in v] \wedge \neg u \in v) \vee$$
$$\vee (x = [v = w] \wedge\ v = w) \vee$$
$$\vee (x = [\neg v = w] \wedge\ \neg v = w) \vee$$
$$\vee (x = \dot{\neg}\dot{\neg} v \wedge\ v \in a) \vee$$
$$\vee (x = v \dot{\wedge} w \wedge\ v \in a \wedge w \in a) \vee$$
$$\vee (x = \dot{\neg}(v \dot{\wedge} w) \wedge\ (\dot{\neg} v \in a \vee \dot{\neg} w \in a)) \vee$$

$$\vee\,(x = \dot{\forall} v \wedge \forall z(v \cdot z \in a)) \vee$$
$$\vee\,(x = \dot{\neg}\dot{\forall} v \wedge \exists z(\dot{\neg} v \cdot z \in a))]$$

Clearly $\Psi(a) := \{x \mid \mathcal{T}(x,a)\}$ is $\subseteq$-monotone in a and is predicative: it receives type 2 once we assign type 0 to u, z, type 1 to x, v, w, type 2 to a.

Definition 3.7

$$Cl_T(a) := \forall x(\mathcal{T}(x,a) \to x \in a)$$
$$W := \{x \mid \forall a(Cl_T(a) \to x \in a)\}$$

The fixed point lemma 3.5 immediately implies:

Proposition 3.8 NFI *proves:*

1. $\exists y(y = W)$;
2. $\forall a(\mathcal{T}(a,W) \to a \in W)$;
3. $Cl_T(a) \to W \subseteq a$.

Remark 4 The interpretation of the truth predicate thus requires *an inductive definition over the universe which still yields a set, i.e. an object of the universe*. This makes essential use of the *peculiar impredicative features* of NFI. If we should try to carry out such definition over the standard set theoretic universe of ZFC, T would result in a *proper* class, and hence we should be forced to apply an impredicative theory of classes à la Morse-Kelley.

Definition 3.9 We inductively (and simultaneously) specify a translation $(-) \mapsto (-)^\nu$ of terms and formulas of SFT into NF:

$$x^\nu := x$$
$$pred(t,s)^\nu = [s^\nu \in t^\nu]$$
$$T(t)^\nu = t^\nu \in W$$
$$tr(t)^\nu = [t^\nu \in W]$$
$$(t = s)^\nu = (t^\nu = s^\nu)$$
$$(id(t,s))^\nu = [t^\nu = s^\nu]$$
$$(A \wedge B)^\nu = A^\nu \wedge B^\nu$$
$$and(t,s)^\nu = t^\nu \dot{\wedge} s^\nu$$
$$(\neg A)^\nu = \neg A^\nu$$
$$neg(t)^\nu = \dot{\neg} t^\nu$$
$$(\forall x A)^\nu = (\forall x A^\nu)$$
$$[x|A]^\nu = \{x|A^\nu\}$$
$$\mu y.r(y,\vec{x})^\nu = \{u|\forall z(r^\nu(z,\vec{x}) \subseteq z \to u \in z)\}$$

$$all(t)^\nu = \dot{\forall} t^\nu$$

If $\vec{x} := x_1, \ldots, x_n$, $\vec{y} := y_1, \ldots, y_n$, $\vec{x} \subseteq \vec{y}$ means $x_1 \subseteq y_1, \ldots, x_n \subseteq y_n$. A term $t(\vec{x})$ is monotone if $\vec{x} \subseteq \vec{y}$ implies $t(\vec{x}) \subseteq t(\vec{y})$.

Lemma 3.10

(i) If A (t) is a stratified formula (term) of SFT, then A^ν (t^ν) is a stratified formula (term) of NF, under the same type assignment to variables and terms of A. Furthermore, if $t(\vec{x})$ is stratified operative in $\vec{x}$, then $t(\vec{x})^\nu$ is stratified monotone in $\vec{x}$.

(ii) If A is a stratified formula of SFT, NF proves:

$$A^\nu[x := u] \leftrightarrow [A^\nu[x := u]] \in W \tag{19.9}$$

Proof As to (i), proceed by simultaneous induction on the definition of *term, term operative in a given parameter* and *formula.*

If t is a variable, the claim is trivial. If $A := T(t)$ is stratified, then so is t; hence by IH t^ν is stratified as well as $t^\nu \in W$.

Let $A := t = s$ be stratified. Then so are t^ν and s^ν. But this implies that $(t = s)^\nu \equiv t^\nu = s^\nu$ is stratified.

Let $t^\nu := (pred(s,r))^\nu$ be stratified. Then $t^\nu = [r^\nu \in s^\nu]$ is stratified too, since by IH the type assignment is preserved by IH. The cases where t (A) is built up by means of *and, id, all, tr* ($\wedge, \forall$) are straightforward by IH.

If $t := [x|A]$, then A^ν is stratified, whence $t := \{x|A^\nu\}$ is stratified.

Let $t := \mu y.r(y,\vec{x})$ be stratified operative in $\vec{x}$. Then $r(y,\vec{x})$ is stratified and operative in $y, \vec{x}$. Hence $(\mu y r(y,\vec{x}))^\nu = \{u|\forall z(r^\nu(z,\vec{x}) \subseteq z \to u \in z)\}$ is stratified and by IH $r^\nu(y,\vec{x})$ is monotone in y and $\vec{x}$. Hence $(\mu y r(y,\vec{x}))^\nu$ is monotone in $\vec{x}$, i.e., if $\vec{a} \subseteq \vec{b}$

$$(\mu y r(y,\vec{a}))^\nu \subseteq (\mu y r(y,\vec{b}))^\nu.$$

If $t(\vec{x}) := pred(r,s)$ is stratified operative in $\vec{x}$, r has type one greater than the type of s and r is operative in $\vec{x}$. Hence by IH r^ν is stratified monotone in $\vec{x}$ with type one greater than the type of s^ν, which is also stratified. It follows by definition of the ν-translation, the property 19.8 and Quine's pairing that $t(\vec{x})^\nu$ is stratified and monotone in $\vec{x}$.

Let us check the case where $t(\vec{x})$ is stratified operative in $\vec{x}$ and

$$t(\vec{x}) = id(s(\vec{x}), r(\vec{x}))$$

Then we have to show that, if $\vec{x} \subseteq \vec{y}$, $t(\vec{x})^\nu \subseteq t(\vec{y})^\nu$. By IH we have

$$\vec{x} \subseteq \vec{y} \to s(\vec{x}) \subseteq s(\vec{y})$$
$$\vec{x} \subseteq \vec{y} \to r(\vec{x}) \subseteq r(\vec{y})$$

The conclusion follows by 19.8, definition of id and Quine's pairing. The remaining cases when t is built up by means of *neg, all* are similar.

(ii): by induction on A, applying part (i) and proposition 3.8 on W. We only consider the case of the universal quantifier. Then by applying stratified comprehension in the last step:

$$\begin{aligned}[\forall x A]^v \in W &\leftrightarrow \forall u([u \in \{x|A^v\}] \in W)\\ &\leftrightarrow \forall u(u \in \{x|A^v\})\\ &\leftrightarrow \forall u A^v[x := u] \equiv (\forall x A)^v\end{aligned}$$

□

The v-translation induces an interpretation into NF:

Theorem 3.11 If $\mathsf{SFT} \vdash A$, then $\mathsf{NF} \vdash A^v$.

Proof It is enough to prove the v-translation of the SFT-axioms. We repeatedly use proposition 3.8 and the independence (or injectivity) of the chosen representation for the logical and descriptive symbols (see definition 3.6).

(i) T is total on predication. Consider e.g. the v-translation of

$$T(pred(y,x)) \vee T(neg(pred(y,x))$$

This amounts to verify

$$([x \in y] \in W \leftrightarrow x \in y) \wedge ([\neg x \in y] \in W \leftrightarrow \neg(x \in y)), \qquad (19.10)$$

which in turn follows from the second and third clauses of the inductive definition of W. By 19.10 also

$$[x \in y] \in W \vee [\neg x \in y] \in W \qquad (19.11)$$

for every x, y. Hence if we choose $y := W$, we obtain:

$$[x \in W] \in W \vee [\neg x \in W] \in W \qquad (19.12)$$

But 19.12 implies the v-translation of the T-axioms involving tr. The verification of the extant cases ($=$, $\wedge$, $\forall$) is also routine.

(ii) T-consistency. Then we must prove the corresponding v-translation, i.e.

$$\neg(x \in W \wedge (\dot{\neg}x) \in W)$$

Choose $\psi(x) := \neg((\dot{\neg}x) \in W)$. Then $\{x|\psi(x)\}$ is a set in NFI and it is easy to check:

$$\forall x(\mathcal{T}(x, \{x|\psi(x)\}) \rightarrow \psi(x))$$

The conclusion is a consequence of proposition 3.8, item 3.

(iii) Stratified β-conversion: we want, if A is stratified,

$$(T(pred([x|A], u)))^v \leftrightarrow (T[A[x := u]])^v$$

By definition of W with proposition 3.8, stratified comprehension and lemma 3.10, we have:

$$\begin{aligned}(\dot{T}(pred([x|A], u))^v &\leftrightarrow [u \in \{x|A^v\}] \in W)\\ &\leftrightarrow u \in \{x|A^v\}\\ &\leftrightarrow A^v[x := u]\\ &\leftrightarrow [A^v[x := u]] \in W\\ &\leftrightarrow (T[A[x := u]])^v\end{aligned}$$

The remaining β conversion schema is similar.

(iv) Self-reference: let $t(y, \vec{x})$ be stratified operative in $y, \vec{x}$. Then by the lemma 3.10, $t^v(y, \vec{x})$ is stratified monotone in y, $\vec{x}$. Hence $\mu y t(y, \vec{x})^v = \{u|\forall z(t^v(z, \vec{x}) \subseteq z \rightarrow u \in z)\}$ satisfies the due fixed point equation by lemma 3.5.

(v) μ-extensionality: straightforward.

(vi) Logical operators are injective: the v-translation of the corresponding axioms is sound, simply because the logical operators act as ordered sequence operators, built upon Quine's ordered pair.

(vii) Logical operators are not surjective: in fact there are objects (e.g.$(4, a)$) which differ from $\dot{\neg}x$, $x \dot{\wedge} y$, $\dot{\forall}$ and $\dot{\in} xy$ (use lemma 3.2). The images of the logical operators are trivially disjoint (we use distinct Fregean numbers as labels), and there are objects which are not P-forms, e.g. any ordered pair $(\emptyset, a)$.

□

Remark 5 Observe that the full strength of stratified comprehension is exploited in interpreting predication. Once predication is restricted to loosely stratified (or predicative) formulas, the resulting version of SFT becomes reducible to a consistent subsystem of NF.

Remark 6 Clearly we can try to reverse the embedding: one can define $x \dot{\in} y$ as $T(pred(y, x))$. Then by the theorem 1.8 it is possible to prove the translation of the NF-stratified comprehension in SFT. However, there is by no means guarantee that SFT proves the translation of the extensionality axiom, and this raises the problem of the consistency strength of SFT.

19.4 Conclusion: Stratified Truth?

Let us try to assess some limits of the theory.

Why stratification? On one hand, that there is a type raising when we move from the mere claim of A to the claim of $T[A]$, can on intuitive grounds be conceded (at least according to the present author). And this is a good reason to pursue the

typed theories of truth[9]. On the other hand, our practice with natural language tends to support the idea that we have to deal with *the* truth predicate, without any further type qualification; hence types ought to be left implicit or possibly avoided. Stratification can be regarded as a way to make both sides coexist. Of course, the awkward aspect is that we do appeal to a theory, which has certain unnatural features. Moreover, the consistency of SFT relies in its full strength upon a discipline – stratification – which is not fully understood, as shown by the yet unsolved problem whether NF be consistent or not.

An additional unsatisfactory point is that the syntactical apparatus of SFT is, at the present stage of formalization, rather complex, and the stratification device is not so transparent as the corresponding explicit typed versions of truth.

Nevertheless, though type-theoretic in essence, SFT allows limited, yet non-trivial forms of self-reference, which are based after all on a semantical construction. And these limitations are apparently essential, in order to preserve consistency. In contrast with usual formal theories of truth, a distinctive feature of SFT is that it allows forms of *direct* self-reference (to make this clear with an example from recursion theory, the second recursion theorem instantiates indirect self-reference, while the first recursion theorem typically supports direct self-reference).

A positive interesting point might be that the compositional axioms of truth receive an unrestricted formalization in SFT, and the truth predicate believes that they are true (in sharp contrast, say, with Kripke-like systems). If we compare SFT with other strong axiomatic systems of truth, we must stress that a high degree of impredicativity is gained. The ground for it is the idea that the basic membership relation is well-defined and given, as made clear by the axiom of well-definedness for truth, and by the semantical clauses in NF governing the operator for inductively defining truth.

As to the relation with the literature, Holmes 2001 explores the possibility that formal semantics is expressed in Quine's NFU, i.e. NFU with urelemente. In particular he shows that the reason why Tarski's argument fails, is not the undefinability of truth, but that the quotation operation becomes type-raising, causing the predicate needed for the 'Tarski sentence' to be unstratified and blocking diagonalization. Now, as already seen, something related happens in our case: the operation for encoding formulas of the form $T(x)$, $\neg T(x)$ is also type raising, and this forbids a form of the Liar leading to inconsistency. Of course, this is at present only a surface analogy. Indeed, a comparative look at Holmes 2001 makes clear a specific limitation of SFT: its truth predicate T is not intended for metamathematical applications, as it is not defined on the inductively defined set of (codes of) sentences of the given *formal language*, say, of NF itself. T can only be applied to objects of the intended universe, which stand for propositions, whatever this means. In other words, the truth notion of SFT is an ontological notion, and is alien to standard semantical arguments, which make use of truth or satisfaction for inductively testing some form of (partial) soundness of the provability tools.

[9] For a thorough critical discussion of the distinction between typed and type-free theories of truth, we send the reader to Halbach 2011, especially part II, and Chaps. 10–11 in part III.

Just as the study of axiomatic theories of truth over standard set theory ZFC has been recently developed (see Fujimoto 2012), the investigation of axiomatic notions of truth over non-standard set theories like NF might be the next reasonable step to the present work.

References

Aczel, P. (1980). Frege structures and the notions of proposition, truth and set. In J. Barwise, H. J. Keisler, K. Kunen (Eds.), *The Kleene Symposium* (pp. 31–59). Amsterdam: North Holland.

Cantini, A. (2004). On a Russellian paradox about propositions and truth. In G. Link (Ed.), *One Hundred Years of Russell's Paradox. Mathematics, logic and philosophy* (pp. 259–284) Berlin: Walter de Gruyter.

Crabbè, M. (1982). On the consistency of an impredicative subsystem of Quine's NF. *The Journal of Symbolic Logic, 47*, 131–136.

Fujimoto, K. (2012). Classes and truths in set theory. *Annals of Pure and Applied Logic, 163*(11), 1484–1523.

Halbach, V. (2011). *Axiomatic theories of truth*. Cambridge: Cambridge University Press.

Holmes, M. R. (1995). The equivalence of NF-style set theories with "tangled" type theories: the construction of ω–models of predicative NF (and more). *The Journal of Symbolic Logic, 60*, 178–190.

Holmes, M. R. (2001). Tarski's theorem and NFU. In C. A. Anderson & M. Zelëny (Eds.), *Logic, meaning and computation. Esssays in memory of Alonzo Church, Synthese Library* (Vol. 305, pp. 469–478). Dordrecht: Kluwer Academic Publishers.

Rosser, J. B. (1953). *Logic for mathematicians*. New York: Mc Graw-Hill.

Specker, E. (1953). The axiom of choice in Quine's New Foundations for Mathematical Logic. *Proceedings of the National Academy of Sciences of the U.S.A., 39*, 972–975. Cambridge: Cambridge University Press.

Russell, B. (1903). *The Principles of Mathematics*. London (reprinted by Routledge, London 1997).