10. Exegetically, I am not sure that Kant would agree that the "same" proposition can be known empirically or a priori. It may be internal to a proposition that it is empirical, say, or a priori. For Kant sometimes says that a proposition changes when it becomes known stated a priori. But, first, this is a conflation that Frege explicitly avoids: in the *Grundlagen* he explicitly says that "a priori" has to do with the justification, not the content, of a proposition. Second, even Kant would have to explain why philosophers have often been confused over whether a proposition is a priori or empirical, if they are not the same propositions even in principle.
11. Am I saying, then, that "9 is the number of planets" is an a priori truth for Frege? Of course not: Frege's deep point is that empirical truths need not be "about" empirical objects – they can be "about" empirical *concepts*.

REFERENCES

Benacerraf, Paul (1981). "Frege: The Last Logicist," *Midwest Studies in Philosophy*, 6: 17–35.

Dauben, Joseph W. (1993). "Review of Joan Weiner's *Frege in Perspective*," *Isis* 84: 618–19.

Demopoulos, William (1995). *Frege's Philosophy of Mathematics*, ed. William Demopoulos (Cambridge, MA: Harvard University Press).

Dummett, Michael (1991). *Frege: Philosophy of Mathematics* (Cambridge, MA: Harvard University Press).

Frege, Gottlob (1959). *The Foundations of Arithmetic*, translated by J. L. Austin (Oxford: Blackwell); republished by Northwestern University Press.

Parsons, Charles (1965). "Frege's Theory of Number," in M. Black (ed.), *Philosophy in America* (Ithaca, NY: Cornell University Press), pp. 180–203.

Parsons, Charles (1979–80). "Mathematical Intuition," *Proceedings of the Aristotelian Society* 80 New Series: 142–68.

Parsons, Charles (1982). "Objects and Logic," *The Monist*, 65: 491–516.

A Theory of Sets and Classes

PENELOPE MADDY

The nature of classes – particularly proper classes, collections "too large" to be sets – is a perennial problem in the philosophy and foundations of set theory. Logicians worry that the unrestricted quantifiers of set theory must range over the collection of all sets, a collection that cannot itself be a set, and hence, a collection that is ill-understood; philosophers puzzle over the existence of properties (such as $x \notin x$) that seem to have no extensions; set theorists ponder heuristic arguments that involve performing operations on the entire universe, V, of sets as if it were a set. Existing theories of sets and classes seem unsatisfactory because their 'proper classes' are either indistinguishable from extra layers of sets or mysterious entities in some perpetual, atemporal process of becoming. In the spirit of Cantor's bold introduction of the completed infinite, we might hope for a theory of sets and classes that both distinguishes plausibly between the two and treats classes as bona fide entities. In the end, this may be too much to ask, but it seems at least the right place to begin.

Several of Charles Parsons' papers have addressed the difficult problem of sets and classes in insightful and influential ways.[1] One central thrust of his treatment has been to emphasize the strong analogy between paradoxes of truth, such as the liar, and paradoxes of classes, such as Russell's paradox. Another has been his clear distinction between sets as mathematical entities, determined combinatorially in a series of stages, and classes as logical entities, determined as extensions of predicates or properties. As Parsons notes, the goals of set theory as a foundation and a branch of mathematics can be met by restricting attention to the mathematical collections, as understood in ZFC. This makes it seem that the paradoxes of classes have been resolved, while the paradoxes of truth have not. But what has really happened is that the mathematical sets have been distinguished from the logical classes, and the emphasis of study has shifted from the latter to the former. The paradoxes of the logical classes remain as stubborn as the paradoxes of truth.

As is well known, Kripke and his successors present an approach to truth that resolves the paradoxes by allowing truth-value gaps.[2] Here I propose an analogous theory of logical classes that resolves their paradoxes by allowing

gaps in the membership relation.[3] As it happens, this approach is something of an anachronism, for it seems that Kripke-like constructions for the case of classes preceded Kripke's own work on truth.[4] The difference between that work and what I propose here is that those theories aimed to supplant typed mathematical sets with untyped logical classes, while my goal is to provide a theory that includes and coordinates both kinds of entities. The earlier constructions could be carried out over any ground model, in particular, over a ground model of set theory, and the resulting structure would have a new class membership relation, distinct from any relation, including set membership, that might be present in the ground model; in contrast, the result of my construction will be a model with one membership relation involving both sets and classes. This seems to me preferable, as, for example, both the set of natural numbers and the class of all infinite collections[5] ought to be members of the class of all infinite collections, members in the same sense that two is a member of the set of natural numbers.

I will be aiming, then, for a theory of sets and classes that characterizes sets as usual, that is, as combinatorially determined in stages, and classes as extensions, capable of such things as self-membership. The construction and accompanying discussion will take up Section I. Section II contains a discussion of the oddities of equality and extensionality, Section III takes up axiomatization and fixed points, and Section IV (given the Fregean inspiration behind much interest in classes) explores the notion of equinumerosity.

I. The Construction

Begin with a first-order language with equality and '$\in$' as its only nonlogical symbol. To this language, add a term-forming operator '$\hat{}$' (called 'hat'), to form terms such as $\hat{x}(x = x)$, $\hat{x}(x \in \emptyset)$, and $\hat{x}(\exists y(x \in \hat{z}(z \notin y)))$.[6] This machinery alone would be enough to handle cases like these three – $\hat{x}(x \in \emptyset)$ could be simulated by assignment of $\emptyset$ to the free variable in $\hat{x}(x \in y)$ – but we make two further additions: for added expressive power, a constant $\overline{V}$ to stand for the class of all sets,[7] and for added simplicity, a constant $\bar{a}$ for every set a. The second addition was apparently standard in the older theories,[8] but it is stronger here, when the intended ground model contains all sets. Though the simplicity gained is great, it could be done without, as will be indicated below.

Define 'formula of $\mathscr{L}$' and 'term of $\mathscr{L}$' simultaneously, as follows[9]:

(i) All constants and variables are terms.
(ii) If t and t' are terms, then $t = t'$ and $t \in t'$ are formulas.
(iii) If ϕ and ψ are formulas, and x is a variable, then $\sim\phi$, $\phi \wedge \psi$, and $\forall x\phi$ are formulas.

(iv) If ϕ is a formula, and x is among the free variables of ϕ, then $\hat{x}\phi$ is a term.

The collection of all terms, T, is the union of S, the collection of all set constants, C, the collection of all $\mathscr{L}$-terms of the form $\hat{x}\phi$, and $\{\overline{V}\}$. C^* is the collection of closed class terms in C; T^* is the collection of all closed terms.[10]

Semantically, we think of $\mathscr{L}$ as a partly interpreted language: the intended domain includes all sets, the constant $\bar{a}$ stands for the set a, the constant $\overline{V}$ stands for a class whose extension includes all sets and whose antiextension includes all classes. The only variable part of the interpretation of $\mathscr{L}$ is the extensions and antiextensions assigned to class terms in C^*.

DEFINITION: $\mathfrak{S} = \{(t, t^+_{\mathfrak{S}}, t^-_{\mathfrak{S}}) \mid t \in C^*\}$ is an $\mathscr{L}$-structure iff for all $t \in C^*$, $t^+_{\mathfrak{S}} \subseteq T^*$ and $t^-_{\mathfrak{S}} \subseteq T^*$, and $t^+_{\mathfrak{S}} \cap t^-_{\mathfrak{S}} = \emptyset$.

The idea is that $t^+_{\mathfrak{S}}$ and $t^-_{\mathfrak{S}}$ represent the extension and the antiextension of the class term t; sets and classes are represented in $t^+_{\mathfrak{S}}$ and $t^-_{\mathfrak{S}}$ by their terms. We leave open the possibility that $t^+_{\mathfrak{S}} \cup t^-_{\mathfrak{S}} \neq T^*$, that is, the possibility that some sets or classes will appear in neither the extension nor the antiextension of t.[11]

Because $\mathfrak{S}$ is a partial interpretation in this sense, we must distinguish three possibilities for a given sentence σ: $\mathfrak{S} \models \sigma$ ($\mathfrak{S}$ thinks σ is true), $\mathfrak{S} \not\models \sigma$ ($\mathfrak{S}$ thinks σ is false), and $\mathfrak{S} \overset{?}{\models} \sigma$ ($\mathfrak{S}$ has no opinion about σ). The 'thinks' relation will be defined only for sentences, beginning with the atomic case:

DEFINITION: If σ is of the form $t \in t'$, for t, t' in T^*, then $\mathfrak{S} \models \sigma$ iff

(i) t is $\bar{a}$, t' is $\bar{b}$, and $a \in b$, or
(ii) $t \in S$ and t' is $\overline{V}$, or
(iii) $t' \in C^*$ and $t \in (t')^+_{\mathfrak{S}}$.

$\mathfrak{S} \not\models \sigma$ iff

(i) t is $\bar{a}$, t' is $\bar{b}$, and $a \notin b$, or
(ii) t is $\overline{V}$ and $t' \in S \cup \{\overline{V}\}$, or
(iii) $t \in C^*$ and $t' \in S \cup \{\overline{V}\}$, or
(iv) $t' \in C^*$ and $t \in (t')^-_{\mathfrak{S}}$.

DEFINITION: If σ is of the form $t = t'$ for t, t' in T^*, then $\mathfrak{S} \models \sigma$ iff t and t' are the same term, and $\mathfrak{S} \not\models \sigma$ iff t and t' are different terms.[12]

Finally, with these interpretations of the atomic sentences, truth and falsity for complex sentences are defined using the strong Kleene rules.

DEFINITION: For $\mathscr{L}$-sentences σ and τ,

$\mathfrak{C} \models \sim\sigma$ iff $\mathfrak{C} \not\models \sigma$; $\mathfrak{C} \not\models \sim\sigma$ iff $\mathfrak{C} \models \sigma$.
$\mathfrak{C} \models \sigma \wedge \tau$ iff $\mathfrak{C} \models \sigma$ and $\mathfrak{C} \models \tau$; $\mathfrak{C} \not\models \sigma \wedge \tau$ iff $\mathfrak{C} \not\models \sigma$ or $\mathfrak{C} \not\models \tau$.
$\mathfrak{C} \models \forall x\phi$ iff for all $t \in T^*$, $\mathfrak{C} \models \phi(t/x)$.
$\mathfrak{C} \not\models \forall x\phi$ iff for some $t \in T^*$, $\mathfrak{C} \not\models \phi(t/x)$.

Then '$\vee$', '$\supset$', '$\equiv$', and '$\exists$' can be defined from these in the usual ways.[13]

It is easy to see that these notions are monotonic; that is, if we define

DEFINITION: If $\mathfrak{C}$ and $\mathfrak{C}'$ are two $\mathscr{L}$-structures, then $\mathfrak{C} \sqsubseteq \mathfrak{C}'$ iff for all $t \in C^*$, $t^+_{\mathfrak{C}} \subseteq t^+_{\mathfrak{C}'}$ and $t^-_{\mathfrak{C}} \subseteq t^+_{\mathfrak{C}'}$.

Then it is an easy induction on the complexity of formulas to show that

PROPOSITION: If $\mathfrak{C} \sqsubseteq \mathfrak{C}'$, then for all $\mathscr{L}$-sentences σ, if $\mathfrak{C} \models \sigma$, then $\mathfrak{C}' \models \sigma$, and if $\mathfrak{C} \not\models \sigma$, then $\mathfrak{C}' \not\models \sigma$.

In other words, once a sentence is decided, adding more elements to the extensions and antiextensions of classes does not disturb that decision.

This simple semantic theory is bought at the expense of a fairly extravagant syntax, that is, at the expense of 'proper class many' constants. As mentioned earlier, this could be avoided by sticking to a language without set constants and allowing assignments to free variables in the usual Tarskian fashion. The semantics for the language would then include interpretations (i.e., extensions and antiextensions) for terms coupled with assignments of constants to free variables, that is, for items of the form $(t, \bar{p})$, where $\bar{p}$ assigns members of $S \cup \{\overline{V}\}$ to the free variables of t. The drawback is the complexity of the definition of satisfaction. When an assignment s to the free variables of ϕ assigns one of these $(t, \bar{p})$s to a free variable that itself occurs within a class term, the scheme becomes notationally complex. In general, to determine whether or not $\mathfrak{C}$ thinks ϕ at s, we must first adjust the free variables of any terms that s assigns so that they do not clash with free variables of ϕ or each other; second, substitute the adjusted t-part of any assigned $(t, \bar{p})$s into the relevant places in ϕ; and third, adjust s to make the assignments dictated by the appropriate $\bar{p}$s.[14] To avoid all this, I prefer to complicate the syntax.

With this machinery in place, we construct a sequence of $\mathscr{L}$-structures, $\mathfrak{C}_0$, $\mathfrak{C}_1$, $\mathfrak{C}_2$, as follows:

$$\mathfrak{C}_0 = \{(\hat{x}\phi, \hat{x}\phi_0^+, \hat{x}\phi_0^-) \mid \hat{x}\phi \in C^*\}, \text{ where } \hat{x}\phi_0^+ = \hat{x}\phi_0^- = \emptyset.$$
$$\mathfrak{C}_{\alpha+1} = \{(\hat{x}\phi, \hat{x}\phi_{\alpha+1}^+, \hat{x}\phi_{\alpha+1}^-) \mid \hat{x}\phi \in C^*\}, \text{ where}$$
$$\hat{x}\phi_{\alpha+1}^+ = \{t \in T^* \mid \mathfrak{C}_\alpha \models \phi(t/x)\} \text{ and } \hat{x}\phi_{\alpha+1}^- = \{t \in T^* \mid \mathfrak{C}_\alpha \not\models \phi(t/x)\}.$$

For λ a limit ordinal,

$$\mathfrak{C}_\lambda = \{(\hat{x}\phi, \hat{x}\phi_\lambda^+, \hat{x}\phi_\lambda^-) \mid \hat{x}\phi \in C^*\}, \text{ where}$$
$$\hat{x}\phi_\lambda^+ = \bigcup_{\alpha<\lambda} \hat{x}\phi_\alpha^+ \text{ and } \hat{x}\phi_\lambda^- = \bigcup_{\alpha<\lambda} \hat{x}\phi_\alpha^-.$$

Finally, we define an $\mathscr{L}$-structure U (for 'universe'):

$$U = \{(\hat{x}\phi, \hat{x}\phi^+, \hat{x}\phi^-) \mid \hat{x}\phi \in C^*),$$

where

$$\hat{x}\phi^+ = \bigcup_{\alpha\in\text{Ord}} \hat{x}\phi_\alpha^+ \quad \text{and} \quad \hat{x}\phi^- = \bigcup_{\alpha\in\text{Ord}} \hat{x}\phi_\alpha^-.$$

This U and the theory it generates are what interest us. By monotonicity, whatever becomes true (or false) at one of the $\mathfrak{C}_\alpha$s remains true (or false) in U. To take one easy example, $\mathfrak{C}_0 \models \bar{\emptyset} \in \overline{\{\emptyset\}}$, and so, $\bar{\emptyset} \in \hat{x}(x \in \overline{\{\emptyset\}})_1^+$; thus, $\mathfrak{C}_1 \models \bar{\emptyset} \in \hat{x}(x \in \overline{\{\emptyset\}})$, which in turn implies that $U \models \bar{\emptyset} \in \hat{x}(x \in \overline{\{\emptyset\}})$.

The classes of U are strictly monadic, but 'ordered tuples' can be defined in various ways. The simplest plays on our strict definition of '='.

DEFINITION: For $t, t' \in T^*$, '(t, t')' is $\hat{z}(z = t \vee z = t')$.[15]

Since $\hat{z}\,(z = t \vee z = t')$ is not the same symbol as $\hat{z}\,(z = t' \vee z = t)$, it is easily established that

PROPOSITION: For $t,\ t' \in T^*$, $U \models ((t, t') = (u, u'))$ iff $U \models (t = t' \wedge u = u')$.

Notice that these 'ordered classes' are total – that is, every element of T^* is in either the extension or the antiextension – so it follows that $U \not\models ((t, t') = (u, u'))$ iff $U \not\models (t = t' \wedge u = u')$. The force of 'ordered n-tuples' can then be recovered as usual – $(t, t', t'') = ((t, t'), t'')$ – and the following useful notation introduced.[16]

DEFINITION: If $x_0, x_1, \ldots$ are among the free variables of ϕ, then $\hat{x}_0 \cdots \hat{x}_n\phi$ abbreviates $\hat{z}(\exists x_0 \cdots \exists x_n(z = (x_0, x_1, \ldots x_n) \wedge \phi))$, where z is the first variable (in some canonical listing) that does not appear in ϕ.

Continuing the earlier example, $\mathfrak{C}_0 \models \bar{\emptyset} \in \overline{\{\emptyset\}}$, so, $\mathfrak{C}_0 \models \exists x \exists y((\bar{\emptyset}, \overline{\{\emptyset\}}) = (x, y) \wedge x \in y)$, which means that $(\bar{\emptyset}, \overline{\{\emptyset\}}) \in \hat{z}(\exists x \exists y(z = (x, y) \wedge x \in y))_1^+$, and thus that $\mathfrak{C}_1 \models ((\bar{\emptyset}, \overline{\{\emptyset\}}) \in \hat{x}\,\hat{y}(x \in y))$. By monotonicity, U agrees.

More substantively, U underwrites our original intuition that $\hat{x}$ (x is infinite) is self-membered. For a sketch of the proof,[17] call a collection, x, 'relational' iff $\forall y(y \in x \supset \exists u \exists v(y = (u, v)))$. A non-empty relational collection will be a class, not a set, because it has classes as members. We can then use the standard methods to define what it is to be a domain or range of a relational class[18]

and what it is for a relational class to be functional or one-to-one. For any set function f, let f^* be $\hat{x}(\exists y \exists z(\langle y, z\rangle \in \overline{f} \wedge x = (y, z)))$, where $\langle y, z\rangle \in \overline{f}$ is the usual statement that the Kuratowski ordered pair of y and z is in $\overline{f}$, but with y, z, and all quantifiers relativized to $\overline{V}$. Then it is not hard (though somewhat tedious) to see that

PROPOSITION: For any sets a and b,

(1) if $\langle a, b\rangle \in f$, then $\mathfrak{C}_1 \models (\bar{a}, \bar{b}) \in f^*$;
(2) if $\langle a, b\rangle \notin f$, then $\mathfrak{C}_1 \not\models (\bar{a}, \bar{b}) \in f^*$;
(3) $\mathfrak{C}_1 \models \forall x(x \in f^* \vee x \notin f^*)$ (i.e., 'f^* is total').

In general, f^* will behave as a class surrogate for f in $\mathfrak{C}_1$.

We can then define 'infinite' in a familiar way.

DEFINITION: 'x is infinite' abbreviates $\exists f$('f is functional' $\wedge$ 'f is one-to-one' $\wedge$ '$\overline{\omega}$ is a domain of f' $\wedge$ 'x contains a range of f').

For any $n \in \omega$, let $n^* = \{n+1, n+2, \ldots\}$. This set is obviously infinite; using f^*, where f is a set function that maps ω one-to-one into n^*, we can easily show that

PROPOSITION: For all $n \in \omega$, $\mathfrak{C}_2 \models \overline{n^*} \in \hat{x}$ (x is infinite).

Let $\phi(y, z)$ be the formula $y \in \overline{\omega} \wedge z \in \overline{\wp(\omega)} \wedge \forall u(u \in \overline{V} \supset (u \in z \equiv u \in \overline{\omega} \wedge y \in u))$. Then,

PROPOSITION: For all $n \in \omega$, $\mathfrak{C}_1 \models (\bar{n}, \overline{n^*}) \in \hat{y}\hat{z}\phi$.

$\mathfrak{C}_1$ also thinks that $\hat{y}\hat{z}\phi$ is functional and one-to-one, and that ω is a domain; by the preceding proposition, $\mathfrak{C}_2$ thinks that $\hat{x}$ (x is infinite) contains a range. It follows that

Theorem. $\mathfrak{C}_3 \models \hat{x}$ (x is infinite) $\in \hat{x}$ (x is infinite).

Finally, as expected, there are gaps in the membership relation of U. The famous example behaves as it should.

PROPOSITION: $U \overset{?}{\models} \hat{x}(x \notin x) \in \hat{x}(x \notin x)$.

Proof: If $\hat{x}(x \notin x)$ were in $[\hat{x}(x \notin x)]^+$, then it would have to enter at some $\hat{x}(x \notin x)^+_\alpha$. Because α cannot be a limit, it must be of the form $\beta + 1$. But then $\mathfrak{C}_\beta \models \hat{x}(x \notin x) \notin \hat{x}(x \notin x)$, and so, $\mathfrak{C}_\alpha \models \hat{x}(x \notin x) \in \hat{x}(x \notin x)$ (because $\hat{x}(x \notin x) \in \hat{x}(x \notin x)^+_\alpha$) and $\mathfrak{C}_\alpha \models \hat{x}(x \notin x) \notin \hat{x}(x \notin x)$ (because $\mathfrak{C}_\beta \sqsubseteq \mathfrak{C}_\alpha$). Contradiction. Similarly, $\hat{x}(x \notin x) \notin \hat{x}(x \notin x)^-$. ■

II. Equality and Extensionality

In a context with membership gaps, extensionality can be formulated as follows: If collections A and B have the same extension and the same antiextension, then $A = B$. (Essentially, we have taken the extension of a set, a, to consist of its members, and its antiextension to consist of everything else (i.e., if we represent it by $T^* - \{\bar{b} \mid b \in a\}$).] Clearly, our definition of '$=$' is not extensional: For example, the terms $\bar{a}$ and $\hat{x}(x \in \bar{a})$ have the same extension and antiextension, but they are not identical. This outcome is perhaps not unwelcome; after all, one of our motivational intuitions was the conviction that sets and classes are entities of two quite different kinds. As a practical matter, identifying classes with coextensional and antiextensional sets would have the result that sets are not total – for example, it would be indeterminate whether $\hat{z}(z \in \bar{a} \wedge (\hat{x}(x \notin x) \in \hat{x}(x \notin x)))$ is a member of $\overline{\{a\}}$ – a decidedly unwelcome result.

Another form of identity relation that might seem attractive is the following natural notion.

DEFINITION: For $t, t' \in T$, '$t \simeq t'$' abbreviates $\forall z(z \in t \equiv z \in t')$, where z is the first variable (in some canonical listing) not in t or t'.

This relation does hold (in U) between coextensive sets and classes like $\bar{a}$ and $\hat{x}(x \in \bar{a})$, and between mildly varying classes like $\hat{x}(x \in \bar{a})$ and $\hat{x}(x \in \bar{a} \wedge x \in \bar{a})$, as well. But the oddities of the connective '$\equiv$' in this context produce some surprises. For U to think, for example, that $u \in t \supset u \in t'$, it must think $u \in t$ is false or $u \in t'$ is true. If U should happen to be undecided about $u \in t$, then it must think $u \in t'$, which in turn means that it must think $u \in t$. This means that for U to think $u \in t \equiv u \in t'$, it cannot be undecided about $u \in t$ or $u \in t'$. Indeed '$t \simeq t$' is a way of expressing the fact that t is total; it is easy to check that $U \models t \simeq t$ if and only if $U \models \forall x(x \in t \vee x \notin t)$. So, $U \overset{?}{\models} (\hat{x}(x \notin x) \simeq \hat{x}(x \notin x \wedge x \notin x))$. For that matter, $U \overset{?}{\models} (\hat{x}(x \notin x) \simeq \hat{x}(x \notin x))$!

Of course, if we confine our attention to sets, the trivial definition of '$=$' adopted here will still guarantee that coextensive sets are equal. (If a and b are coextensive, they are in fact the same set, and so, $\bar{a} = \bar{b}$.) But attempted identifications across the set/class boundary are not the only failures of extensionality produced by our definition; classes can be coextensive without coinciding if they are picked out by different terms. To a certain extent, this seems appropriate, because classes are understood as closely tied to the properties that determine them, and coextensive properties are not identified. Still, it seems a bit much to distinguish $\hat{x}(x \in \bar{a})$ from $\hat{y}(y \in \bar{a})$, and some might balk at distinguishing $\hat{z}(z \in x \vee z \in y)$ from $\hat{z}(z \in y \vee z \in x)$.[19] It might be possible to modify my 'trivial' definition of '$=$' to suit various notions of how properties should be individuated, but I will not take up this topic here.

III. Axiomatics and Fixed Points

Of course, we cannot expect to axiomatize all the truths of a structure as rich as U, but we might hope for an axiomatization that provides as much information about sets as ZFC as well as a usefully large body of truths about classes. Unfortunately, the nonclassical context created by the membership gaps is a serious obstacle; for an authoritative discussion, see Feferman (1984). A telling symptom is that we cannot hope for an axiom of the form $x \in \hat{x}\phi \equiv \phi$, because the biconditional is indeterminate if either side is. The general problem is well illustrated by this version of Curry's paradox, drawn from Flagg and Myhill (1987):

PROPOSITION: Any system with the following properties:

(i) $\Gamma \vdash x \in \hat{x}\phi$ iff $\Gamma \vdash \phi$ (Frege's principle)
(ii) $\Gamma \cup \{\phi\} \vdash \psi$ iff $\Gamma \vdash \phi \supset \psi$ (deduction theorem)
(iii) if $\Gamma \vdash \phi$ and $\Gamma \vdash \phi \supset \psi$, then $\Gamma \vdash \psi$ (Modus ponens)

is inconsistent.

Proof: Let σ be any sentence. Then,

$$\hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma) \vdash \hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma).$$

So, by Frege's principle,

$$\hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma) \vdash (\hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma)) \supset \sigma.$$

By modus ponens,

$$\hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma) \vdash \sigma.$$

By the deduction theorem,

$$\vdash [\hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma)] \supset \sigma.$$

By Frege's principle again,

$$\vdash \hat{x}(x \in x \supset \sigma) \in \hat{x}(x \in x \supset \sigma).$$

So, by modus ponens, $\vdash \sigma$. ■

It is hard to know how to proceed without modus ponens and the deduction theorem. Nevertheless, for the case at hand, Myhill has suggested a straightforward infinitary axiomatization. Though the deduction theorem fails, the system is worth considering for the further questions it raises.

Because the comprehension axiom cannot be effectively expressed using the Kleene '$\equiv$', Myhill uses a battery of rules in addition to axioms. Call the following system M.

AXIOMS: For any $a, b \in V$, and any $t \in C^*$,

(i) $\bar{a} \in \bar{b}$ when $a \in b$,
(ii) $\bar{a} \notin \bar{b}$ when $a \notin b$,
(iii) $\bar{a} \in \overline{V}$,
(iv) $t \notin \bar{a}$,
(v) $t \notin \overline{V}$,

for any distinct $t, t' \in T^*$,

(vi) $t = t$,
(vii) $t \neq t'$.

RULES:

(1) $\dfrac{\phi}{\sim\sim\phi}$

(2) $\dfrac{\phi, \psi}{\phi \wedge \psi}$ $\quad \dfrac{\sim\phi}{\sim(\phi \wedge \psi)}$ $\quad \dfrac{\sim\psi}{\sim(\phi \wedge \psi)}$

(3) $\dfrac{\phi(t/x) \text{ for all } t \in T^*}{\forall x\phi}$ $\quad \dfrac{\sim\psi(t/x) \text{ for some } t \in T^*}{\sim\forall x\phi}$

(4) $\dfrac{\phi(t/x)}{t \in \hat{x}\phi}$ $\quad \dfrac{\sim\phi(t/x)}{t \notin \hat{x}\phi}$

We write $\vdash_M \sigma$ when σ can be derived in this system. Notice that the set constants play an ineliminable role here, for the first time.

M is a success if the same sentences are provable as are true, that is, if for all σ, $\vdash_M \sigma$ if and only if $U \models \sigma$ (and $\vdash_M \sim\sigma$ iff $U \not\models \sigma$). To begin, it is fairly easy to see (by a double induction, first on α, then on the complexity of σ) that if $\mathfrak{S}_\alpha \models \sigma$, for some α, then $\vdash_M \sigma$ (and if $\mathfrak{S}_\alpha \not\models \sigma$, for some α, then $\vdash_M \sim\sigma$). To get from here to the completeness of M requires that $U \models \sigma$ imply the existence of an α such that $\mathfrak{S}_\alpha \models \sigma$ (and that $U \not\models \sigma$ imply the existence of an α such that $\mathfrak{S}_\alpha \not\models \sigma$). This would obviously be true if the construction reached a fixed point, that is, if there were an $\mathfrak{S}_\alpha$ such that $\mathfrak{S}_\alpha = \mathfrak{S}_{\alpha+1}$. As for soundness, we know that $\mathfrak{S}_\alpha \models \sigma$, for some α, implies that $U \models \sigma$ (and $\mathfrak{S}_\alpha \not\models \sigma$, for some α, implies $U \not\models \sigma$). To get from here to soundness, we need to know that $\vdash_M \sigma$ implies the existence of an α such that $\mathfrak{S}_\alpha \models \sigma$ (and $\vdash_M \sim\sigma$ implies the existence of an α such that $\mathfrak{S}_\alpha \not\models \sigma$). Because all the axioms are true in all $\mathfrak{S}_\alpha$s, we could begin by showing that if each antecedent of a rule is true at some $\mathfrak{S}_\alpha$ or other, then there is an α' such that the consequent is true at $\mathfrak{S}_{\alpha'}$. This is obvious for rules 1, 2, and 4, but the quantifiers present a problem: assuming that for every $t \in T^*$, there is an α_t such that $\mathfrak{S}_{\alpha_t} \models \phi(t/x)$, the usual replacement argument[20] does not guarantee the existence of an α larger than all the α_ts (where $\mathfrak{S}_\alpha$ would think $\phi(t/x)$ for all $t \in T^*$, and thus think $\forall x\phi$), because T^* is not a set. In this case, too, the existence of a fixed point would save the day.

So, the existence of a fixed point would imply that the infinitary axiom system M successfully codifies the truths of the structure U. In fact, however, the construction does not reach a fixed point, as can be seen from the following theorem of Tait:

Tait's Theorem. For any formula $\phi(y, x)$, there is a formula $\psi(z)$ such that $\hat{z}\psi^+ = \hat{x}\phi(\hat{z}\psi, x)^+$ and $\hat{z}\psi^- = \hat{x}\phi(\hat{z}\psi, x)^-$. In fact, for all α, $\hat{z}\psi^+_{\alpha+1} = \hat{x}\phi(\hat{z}\psi, x)^+_\alpha$ and $\hat{z}\psi^-_{\alpha+1} = \hat{x}\phi(\hat{z}\psi, x)^-_\alpha$.

Proof: Let A be $\hat{y}\hat{x}\phi(\hat{z}((y, z) \in y), x)$. Let $\psi(z)$ be $(A, z) \in A$, so that $\hat{z}\psi$ is $\hat{z}((A, z) \in A)$. Then, for any $t \in T^*$ and any α, if $t \in \hat{z}\psi^+_{\alpha+1}$, then there is a $\beta < \alpha + 1$ such that $\mathfrak{C}_\beta \models \psi(t)$, that is, such that $\mathfrak{C}_\beta \models (A, t) \in A$. By definition of A, it follows that there is a $\gamma < \beta$ such that $\mathfrak{C}_\gamma \models \phi(\hat{z}((A, z) \in A), t)$, which is just to say that $\mathfrak{C}_\gamma \models \phi(\hat{z}\psi, t)$. Thus, $t \in \hat{x}\phi(\hat{z}\psi, x)^+_{\gamma+1}$, and $\gamma < \beta < \alpha + 1$, so $t \in \hat{x}\phi(\hat{z}\psi, x)^+_\alpha$. Conversely, for any $t \in T^*$ and any α, if $t \in \hat{x}\phi(\hat{z}\psi, x)^+_\alpha$, then there is a $\beta < \alpha$ such that $\mathfrak{C}_\beta \models \phi(\hat{z}\psi, t)$, which is to say that $\mathfrak{C}_\beta \models \phi(\hat{z}((A, z) \in A), t)$. By definition of A, it follows that $\mathfrak{C}_{\beta+1} \models (A, t) \in A$, that is, that $\mathfrak{C}_{\beta+1} \models \psi(t)$. So, $t \in \hat{z}\psi^+_{\beta+2}$, which implies that $t \in \hat{z}\psi^+_{\alpha+1}$. From these two facts, it follows that $\hat{z}\psi^+ = \hat{x}\phi(\hat{z}\psi, x)^+$. Similarly for $\hat{z}\psi^-$. ■

This theorem yields a number of interesting examples, beginning with the one that undermines the possibility of a fixed point.

To see this, suppose that $\phi(y, x)$ is $\forall w(w \in x \supset w \in y)$. If ψ is formed as in the proof of Tait's theorem, then we have $\hat{z}\psi^+ = \hat{x}(\forall w(w \in x \supset w \in \hat{z}\psi)^+$ and $\hat{z}\psi^- = \hat{x}(\forall w(w \in x \supset w \in \hat{z}\psi))^-$. Roughly, then, a collection is in $\hat{z}\psi$ when all its elements are.[21] We see that $\mathfrak{C}_0 \models \forall w(w \in \overline{\emptyset} \supset w \in \hat{z}\psi)$, so $\overline{\emptyset} \in \hat{x}(\forall w(w \in x \supset w \in \hat{z}\psi))^+_1$; applying Tait's theorem, we get $\overline{\emptyset} \in \hat{z}\psi^+_2$. Then, $\mathfrak{C}_2 \models \forall w(w \in \overline{\{\emptyset\}} \supset w \in \hat{z}\psi)$, so $\overline{\{\emptyset\}} \in \hat{x}(\forall w(w \in x \supset w \in \hat{z}\psi))^+_3$, and by Tait's theorem, $\overline{\{\emptyset\}} \in \hat{z}\psi^+_4$. The pattern is clear. By induction on ordinals, it is straightforward to check that:

COROLLARY: For all ordinals α, $\alpha \notin \hat{z}\psi^+_\alpha$.

COROLLARY: For any ordinal α, there is an natural number n such that $\alpha \subseteq \hat{z}\psi^+_{\alpha+n}$ and $\overline{\alpha} \in \hat{z}\psi^+_{(\alpha+n)+2}$.

New ordinals will be entering $\hat{z}\psi$ arbitrarily high up, and so, the construction cannot become constant, and there is no fixed point.[22]

For the record, we can say more about the class $\hat{z}\psi$. We have seen that every ordinal enters $\hat{z}\psi^+$ at some stage not long after its rank, and it is easy to see that every set also will. On the other hand,

PROPOSITION: $\hat{z}\psi^-$ is empty.

Proof: Suppose not, that is, that $\hat{z}\psi^-$ is non-empty; suppose that $t \in T^*$, $t \in \hat{z}\psi^-$, and t enters $\hat{z}\psi^-$ at the first stage at which it is non-empty, say at $\hat{z}\psi^-_\alpha$. Then

α cannot be a limit; let $\alpha = \beta + 1$. By Tait's theorem, we have $t \in \hat{x}(\forall w(w \in x \supset w \in \hat{z}\psi))^-_\beta$, and thus, $\mathfrak{C}_\beta \models \forall w(w \in t \supset w \in \hat{z}\psi)$. For some $t' \in T^*$, $\mathfrak{C}_\beta \models t' \in t \supset t' \in \hat{z}\psi$, and so, $\mathfrak{C}_\beta \models t' \in t$ and $\mathfrak{C}_\beta \models t' \in \hat{z}\psi$. So, $t' \in \hat{z}\psi^-_\beta$, which contradicts the choice of α. ■

Another interesting example is generated by taking $\phi(y, x)$ to be $x \notin y$. Then the class – call it E – generated by Tait's construction includes all collections that are not members of E!

PROPOSITION: $E^+ = E^- = \emptyset$. (So, for all $t \in T^*$, $U \stackrel{?}{\models} t \in E$.)

Proof: If $t \in E^+_\alpha$, then $t \in \hat{x}(x \notin E)^+_\alpha$, so there is a $\beta < \alpha$ such that $\mathfrak{C}_\beta \models t \in E$. But then $\mathfrak{C}_\alpha \models t \in E \wedge t \notin E$. Contradiction. Similarly, for E^-. ■

All class terms (except $\overline{V}$) begin with empty extensions and antiextensions, and E also ends up that way.

In any case, we see that we cannot hope to settle our questions about the soundness and completeness of M by appeal to a fixed point. Still, as it happens, we are now in a position to resolve these questions directly. We have noted that

PROPOSITION: For all sentences σ,

(1) if for some α, $\mathfrak{C}_\alpha \models \sigma$, then $\vdash_M \sigma$; and
(2) if for some α, $\mathfrak{C}_\alpha \not\models \sigma$, then $\vdash_M \sim\sigma$.

We attempted to prove the converse by showing that if each antecedent to a rule is true at some $\mathfrak{C}_\alpha$ or other, then there is a $\mathfrak{C}_{\alpha'}$ where the consequent is also true, and this led to the question of the existence of a fixed point. But the converse can be proved in a slightly subtler way, if we assume (as seems reasonable) that any proof in M will employ α-many applications of M-rules, for some ordinal α.

PROPOSITION: For all sentences σ and all ordinals α,

(1) if σ is M-provable in α steps, then $\mathfrak{C}_\alpha \models \sigma$;
(2) if $\sim\alpha$ is M-provable in α steps, then $\mathfrak{C}_\alpha \not\models \sigma$.

This can be proved by an easy induction on α. So, proofs in M and the construction of U match up as they go along.

But we are interested in what happens at U itself, and here the information gained from Tait's theorem is relevant. So far, we have an example of a class whose extension gains members at arbitrarily late stages. Can we convert this into an example of a ϕ such that (1) for every $t \in T^*$, there is an α_t such that $\mathfrak{C}_{\alpha_t} \models \phi(t/x)$; and (2) for all α, $\mathfrak{C}_\alpha \stackrel{?}{\models} \forall x\phi$? If so, it follows both that M is incomplete and that one of its rules is unsound. For incompleteness, note that for such a ϕ: not $\vdash_M \forall x\phi$ (because this would imply that $\forall x\phi$ is true at some $\mathfrak{C}_\alpha$, which it is not), but $U \models \forall x\phi$. For unsoundness of an M-rule, given

such a ϕ, consider the formula $\forall x\phi \vee y \in E$ and let t be any closed term. Then, $U \models (\forall x\phi \vee y \in E)(t/y)$ (because $U \models \forall x\phi$), but $U \stackrel{?}{\not\models} t \in \hat{y}(\forall x\phi \vee y \in E)$ (because t never enters a $\hat{y}(\forall x\phi \vee y \in E)^+_\alpha$ or a $\hat{y}(\forall x\phi \vee y \in E)^-_\alpha$]. ($M$ would still be sound in the sense that $\vdash_M \sigma$ implies $U \models \sigma$, but it could not be used as a dependable way to move from truths of U to truths of U.) So, the existence of such a ϕ would be of considerable interest.

Knowing that ordinals enter $\hat{z}\psi$ at arbitrarily late stages, we might begin our search for such a ϕ with something like [Ord(x) $\supset$ $x \in \hat{z}\psi$), where 'Ord(x)' is 'x is transitive and $\in$ is connected on x'. The trouble with this thought is that the Kleene interpretation of '$\supset$' would require that Ord(t) be determinately false for any t that is not determinately in $\hat{z}\psi$. But the very empty E is a counterexample.

PROPOSITION: $U \stackrel{?}{\models}$ Ord(E) and $U \stackrel{?}{\models} E \in \hat{z}\psi$.

Proof: Consider just the transitivity clause of Ord(E): $\forall x \forall y(x \in y \wedge y \in E \supset x \in E)$. If $U \models t \in t'$ for some $t, t' \in T^*$, as it often does, then for this pair, U must think $t' \notin E$ or $t \in E$. But as we have just seen, U thinks neither of these things. So, not $U \models$ Ord(E). Similarly, for U to think Ord(E) is false, it must think E is not transitive or $\in$ is not connected on E, both of which require that it think something is definitely in E. But it does not. So not $U \not\models$ Ord(E).

We have already seen that not $U \not\models E \in \hat{z}\psi$. For U to think that $E \in \hat{z}\psi$, there would have to be an α such that $\mathfrak{C}_\alpha \models \forall w(w \in E \supset x \in \hat{z}\psi)$. In other words, for all $t \in T^*$, we would need $\mathfrak{C}_\alpha \models t \notin E \vee t \in \hat{z}\psi$. Now, no $\mathfrak{C}_\alpha$ ever thinks $t \notin E$, and so, we would need: For all $t \in T^*$, $\mathfrak{C}_\alpha \models t \in \hat{z}\psi$. It is easy to check that $\hat{x}(x \notin x)$ is a counterexample, indeed that for all α, $\mathfrak{C}_\alpha \stackrel{?}{\models} \hat{x}(x \notin x) \in \hat{z}\psi$. (We have already seen that $\hat{z}\psi^-$ is empty, so not $\mathfrak{C}_\alpha \not\models \hat{x}(x \notin x) \in \hat{z}\psi$. If, on the other hand, there is an α such that $\mathfrak{C}_\alpha \models \hat{x}(x \notin x) \in \hat{z}\psi$, let α^* be the least. This α^* cannot be a limit, so let $\alpha^* = \beta + 1$. Then $\mathfrak{C}_\beta \models \forall w(w \in \hat{x}(x \in x) \supset w \in \hat{z}\psi)$, so that, in particular, $\mathfrak{C}_\beta \models \hat{x}(x \notin x) \notin \hat{x}(x \notin x) \vee \hat{x}(x \notin x) \in \hat{z}\psi$. $\mathfrak{C}_\beta$ cannot think the first disjunct, and so, it must think the second. But this contradicts the choice of α^*.) So, not $U \models E \in \hat{z}\psi$. ■

So, our first try at the desired ϕ is unsuccessful.

As it happens, the key idea can be revived simply by modifying the formula Ord(x) to make it total. In particular, it can be made explicitly false of all non-ordinals simply by relativizing everything to $\overline{V}$. Let Ord$^*(x)$ be $x \in \overline{V} \wedge$ [Ord$(x)]^{\overline{V}}$. Let $\phi(x)$ be Ord$^*(x) \supset x \in \hat{z}\psi$. Then,

PROPOSITION: (1) For any $t \in T^*$, there is an α_t such that $\mathfrak{C}_{\alpha_t} \models \phi(t/x)$. (2) For all α, $\mathfrak{C}_\alpha \stackrel{?}{\models} \forall x\phi$.

Proof: (1) We have seen that, for any α, there is a natural number n such that $\mathfrak{C}_{\alpha+n} \models \overline{\alpha} \in \hat{z}\psi$. For $t \in C \cup \{\overline{V}\}$ and for t that are set constants $\bar{a}$, where a is not an ordinal, we have $\mathfrak{C}_0 \not\models$ Ord$^*(t)$.

(2) This is just a restatement of our observation that the ordinals enter $\hat{z}\psi^+$ in a cofinal series of stages. ■

As noted earlier, this example compromises both the soundness and the completeness of M.

IV. Equinumerousity and Number

One of the enduring fascinations of classes is the possibility of reviving some approximation of the Fregean theory of numbers. Something along these lines can be accomplished here, though the irritations of the nonclassical context naturally persist.

Begin with a version of Frege's notion of equinumerosity.

DEFINITION: For $t, t' \in T$, '$t \approx t'$' abbreviates

$$\exists z(\forall u \forall v \forall w(((u, v) \in z \wedge (u, w) \in z \supset v = w) \wedge$$
$$((u, v) \in z \wedge (w, v) \in z \supset u = w)) \wedge \forall u((u \in t \supset \exists v(v \in t' \wedge (u, v) \in z)) \wedge$$
$$(u \in t' \supset \exists v(v \in t \wedge (v, u) \in z)))),$$

where z, u, v, w are the first variables (in some canonical listing) that do not appear in t or t'.

If f is a term that truly instantiates this sentence, we write '$t \approx_f t'$'. If a set f is a one-to-one correspondence between sets a and b, in the ordinary sense, recall that there is a class f^* such that $\bar{a} \approx_{f^*} \bar{b}$.

It is perhaps unsurprising that '$\approx$' only captures the intended notion for total collections. If t is not total, and u is in neither t^+ nor t^-, then the second clause of the definition requires that there be a u' in $(t')^+$ such that u is mapped to u'. But since u' is in $(t')^+$, the second clause also requires that there be a u'' in t^+ such that u'' is mapped to u'. But, then, the first clause requires that $u = u''$, so u is in t^+ and u is not in t^+. So, '$\approx$' can never hold between nontotal classes.[23] But it is fairly well behaved for total classes:

PROPOSITION: For all α, for all $t, t', t'' \in T^*$, if $\mathfrak{C}_\alpha \models t \simeq t \wedge t' \simeq t' \wedge t'' \simeq t''$, then

(1) $\mathfrak{C}_\alpha \models t \approx t$.
(2) if $\mathfrak{C}_\alpha \models t \approx t'$, then $\mathfrak{C}_{\alpha+1} \models t' \approx t$.
(3) if $\mathfrak{C}_\alpha \models t \approx t'$ and $\mathfrak{C}_\alpha \models t' \approx t''$, then $\mathfrak{C}_{\alpha+1} \models t \approx t''$.

The proofs are as usual: For example, for (3), if $t \approx_f t'$, and $t' \approx_g t''$, and h is $\hat{x}\hat{y}(\exists z((x, z) \in f \wedge (z, y) \in g))$, then $t \approx_h t''$.

But if the failure of '$\approx$' in the case of nontotal classes is unsurprising, other of its shortcomings are more troublesome. For example, there are very few explicit failures of equinumerousity.

PROPOSITION: For any $t, t' \in T^*$, if $U \models \exists x(x \in t) \wedge \exists x(x \in t')$ and not $U \models t \approx t'$, then $U \stackrel{?}{\models} t \not\approx t'$.

Proof: For U to think $t \approx t'$ is false, it must think that every member of T^* falsifies one of the two clauses in the definition. Consider E. For E to falsify either conjunct of the first clause, it would need to have members, which it does not have. For E to falsify the first conjunct of the second clause, we would need $\exists u(u \in t \wedge \forall v(v \notin t' \vee (u, v) \notin E))$. Because the antiextension of E is empty, this can only happen if the extension of t' is empty. Similarly, the second conjunct of the second clause can only be falsified if t^+ is empty. ■

So, for example, U is undecided about whether or not $\overline{\{\emptyset\}}$ is equinumerous with $\overline{\{\emptyset, \{\emptyset\}\}}$.

Furthermore, if we attempt to define the number of a collection as its equivalence class under equinumerousity, we cannot (given our strict interpretation of '=') recover the fundamental Hume's principle in its usual form: $[t]_\approx = [t']_\approx$ iff $t \approx t'$. We could replace '=' by '$\simeq$', but this runs up against the problem that the equivalence class $[t]_\approx$ is not total, even if t is.

PROPOSITION: For all $t \in T^*$, $U \stackrel{?}{\models} E \approx t$.

Proof: For $U \models E \approx t$, we would need a $z \in T^*$ which, among other things, satisfies $\forall u(u \notin t \vee \exists v(v \in E \wedge (v, u) \in z))$. The second disjunct is never true, and so, we would need $\forall u(u \notin t)$. On the other hand, we also need $\forall u(u \notin E \vee \exists v(v \in t \wedge (u, v) \in z))$. The first disjunct is never true, so for this sentence to be satisfied, t would have to have members. Contradiction. For $U \not\models E \approx t$, we need one clause in the definition of '$\approx$' to be false for any $z \in T^*$. Consider E itself. As noted in the proof of the preceding proposition, for E to falsify the first clause or the first conjunct of the second clause would require E's extension to have members, which it does not have. The only hope is to falsify the final conjunct, that is, to satisfy $\exists u(u \in t \wedge \forall v(v \notin E \vee (v, u) \notin E))$. To do this, E's antiextension would need to have members, which it does not have. ■

Given these constraints, it seems the closest we can get to Hume's principle is something like this: $\hat{x}(x \approx t)^+ = \hat{x}(x \approx t')^+ \neq \emptyset$ iff there is an α such that $\mathfrak{S}_\alpha \models t \approx t'$. ($\Rightarrow$: if $u \in \hat{x}(x \approx t)^+$ and $u \in \hat{x}(x \approx t')^+$, then there are α and β such that $u \in \hat{x}(x \approx t)^+_\alpha$ and $u \in \hat{x}(x \approx t')^+_\beta$. If $\gamma = \max(\alpha, \beta)$, then $\mathfrak{S}_\gamma \models u \approx t \wedge u \approx t'$. So, by an earlier proposition, $\mathfrak{S}_{\gamma+1} \models t \approx t'$. $\Leftarrow$: if $\mathfrak{S}_\alpha \models t \approx t'$, then $\mathfrak{S}_\alpha \models t \approx t$, so $\hat{x}(x \approx t)^+ \neq \emptyset$. So, if $u \in \hat{x}(x \approx t)^+$, then

there is a β such that $u \in \hat{x}(x \approx t)^+_\beta$. If $\gamma = \max(\alpha, \beta)$, then $\mathfrak{S}_\gamma \models t \approx t' \wedge u \approx t$. So, by an earlier proposition, $\mathfrak{S}_{\gamma+1} \models u \approx t'$, and $u \in \hat{x}(x \approx t')^+$. Similarly, if $u \in \hat{x}(x \approx t')^+$, then $u \in \hat{x}(x \approx t)^+$.)

There is another, more concrete, but still quasi-Fregean,[24] approach to number that works a bit better. The general idea, as in contemporary set theory, is to use the von Neumann ordinals as a standard measure of cardinality.[25] Limiting our attention first to the natural numbers, consider the following proposal:

DEFINITION: For $t \in T^*$, let '$[t]_\approx$' be $\hat{x}(x \approx t)$. Let $\mathbb{N}$ be $\hat{x}(\exists\alpha(\alpha \in \overline{\omega} \wedge x = [\alpha]_\approx))$.

Notice that $\mathbb{N}$ is total. (If t is $\overline{V}$ or $\bar{a}$, for some set a, then $t = [\alpha]_\approx$ is false for any $\alpha \in \omega$; if $t \in C^*$, then $t = [\alpha]_\approx$ is determinately true or false for any $\alpha \in \omega$.)

DEFINITION: '$\mathrm{Num}(x, y)$' abbreviates $x \in \mathbb{N} \wedge y \in x$.

0 is $[\overline{\emptyset}]_\approx$, 1 is $[\overline{\{\emptyset\}}]_\approx$, ...

Sxy abbreviates $x \in \mathbb{N} \wedge y \in \mathbb{N} \wedge \exists u \exists v(x = [u]_\approx \wedge y = [v]_\approx \wedge \forall w(w \in u \equiv w \in v \vee w = v))$.

In these terms, our approximation to the Fregean equivalence is a bit better.

PROPOSITION: For all α, and all $t, t', u \in T^*$,

(1) if $\mathfrak{S}_\alpha \models t \approx t' \wedge \mathrm{Num}(u, t)$, then $\mathfrak{S}_{\alpha+2} \models \mathrm{Num}(u, t')$;
(2) if $\mathfrak{S}_\alpha \models \exists x[\mathrm{Num}(x, t) \wedge \mathrm{Num}(x, t')]$, then $\mathfrak{S}_{\alpha+1} \models t \approx t'$.

Proof: (1) if $\mathfrak{S}_\alpha \models t \approx t' \wedge \mathrm{Num}(u, t)$, then there is a $\beta \in \omega$ such that $\mathfrak{S}_\alpha \models u = [\beta]_\approx \wedge t \in u$. So, $\mathfrak{S}_\alpha \models t \approx t' \wedge t \approx \beta$. By an earlier proposition, $\mathfrak{S}_{\alpha+1} \models t' \approx \beta$, so $\mathfrak{S}_{\alpha+2} \models \mathrm{Num}(u, t')$. (2) If $\mathfrak{S}_\alpha \models \mathrm{Num}(u, t) \wedge \mathrm{Num}(u, t')$, then there is a $\beta \in \omega$ such that $\mathfrak{S}_\alpha \models t \approx \beta \wedge t' \approx \beta$. By the same earlier proposition, $\mathfrak{S}_{\alpha+1} \models t \approx t'$. ■

Notice that it is possible to count classes as well as sets using these numbers. (For example, $\hat{x}(x \in x)$ and $\hat{x}(x \approx \overline{\{\emptyset, \{\emptyset\}\}})$ (alias, '2') can be collected into $\hat{y}(y = \hat{x}(x \in x) \vee y = 2)$, and it is then easy to see that $\hat{y}(y = \hat{x}(x \in x) \vee y = 2) \approx \overline{\{\emptyset, \{\emptyset\}\}}$. At the next stage, we get $\hat{y}(y = \hat{x}(x \in x) \vee y = 2) \in 2$, and thus $\mathrm{Num}(2, (y = \hat{x}(x \in x) \vee y = 2))$.)

It is straightforward, though occasionally tedious, to check that the first four Peano axioms become true in the early stages of the construction of U.

PA(1) $0 \in \mathbb{N}$.
PA(2) $\forall x(x \in \mathbb{N} \supset \exists y(y \in \mathbb{N} \wedge S(y, x))$.
PA(3) $\forall x(x \in \mathbb{N} \supset \sim S(0, x))$
PA(4) $\forall x \forall y(x \in \mathbb{N} \wedge y \in \mathbb{N} \wedge \exists z(S(z, x) \wedge S(z, y) \supset x = y)$.

The induction axiom is more delicate. It is false in its usual form – $\forall x((x \subseteq \mathbf{N} \wedge 0 \in x \wedge \forall y(y \in \mathbf{N} \wedge y \in x \supset \exists z(z \in \mathbf{N} \wedge S(z, y) \wedge z \in x))) \supset x = \mathbf{N})$ – because of the strict interpretation of '='. This can be avoided by replacing '=' with '$\simeq$', but the result still fails for nontotal classes like E. (With 'E' in for 'x', none of the antecedents is definitely falsified, nor is the consequent definitely satisfied.) The solution is to trade the axiom for a rule:

$$\frac{t \simeq t, t \subseteq \mathbf{N}, 0 \in \mathbf{N}, \forall y(y \in \mathbf{N} \wedge y \in t \supset \exists z(z \in \mathbf{N} \wedge S(z, y) \wedge z \in t))}{t \simeq \mathbf{N}}$$

More arithmetic can be developed from here, within the limitations of the nonclassical context.

This account of number can obviously be carried into the infinite, using the transfinite ordinals. Let me close this discussion with an few rather fanciful observations. To begin, consider $\hat{x}(\mathrm{Ord}^*(x))$.

PROPOSITION:
(1) $\mathfrak{C}_1 \models \hat{x}(\mathrm{Ord}^*(x)) \simeq \hat{x}(\mathrm{Ord}^*(x))$.
(2) $\mathfrak{C}_2 \models \hat{x}(\mathrm{Ord}^*(x)) \in \hat{x}(\mathrm{Ord}(x))$.

Proof: (1) $\mathfrak{C}_0$ knows which sets are ordinals and which are not, and that only sets are members of $\overline{V}$, to which the quantifiers of Ord^* are relativized. So, $\hat{x}(\mathrm{Ord}^*(x))$ is total at $\mathfrak{C}_1$.

(2) For all $t, t' \in T^*$, $\mathfrak{C}_1 \models t \notin t' \vee t' \notin \hat{x}(\mathrm{Ord}^*(x)) \vee t \in \hat{x}(\mathrm{Ord}^*(x))$ (transitivity) and $\mathfrak{C}_1 \models t \notin \hat{x}(\mathrm{Ord}^*(x)) \vee t' \notin \hat{x}(\mathrm{Ord}^*(x)) \vee t \in t' \vee t = t' \vee t' \in t$ (connectedness of $\in$), and so, $\mathfrak{C}_1 \models \mathrm{Ord}(\hat{x}(\mathrm{Ord}^*(x)))$. Thus, $\mathfrak{C}_2 \models \hat{x}(\mathrm{Ord}^*(x)) \in \hat{x}(\mathrm{Ord}(x))$. ■

Given that $\hat{x}(\mathrm{Ord}^*(x))$ is so ordinal-like, it seems natural to regard $[\hat{x}(\mathrm{Ord}^*(x))]_{\simeq}$ as number-like. To my knowledge, the only previous discussion of such numbers is due to Himmel,[26] who called them 'supernaturals'. There would also be additional 'superordinals', for example, $\hat{y}(y \in \hat{x}(\mathrm{Ord}^*(x)) \vee y = \hat{x}(\mathrm{Ord}^*(x)))$. Some of our earlier problems might be helped by iterating the construction through these superordinal stages, though new superordinals would still be entering Tait's class at each superordinal stage.

This is only the bare beginning of an exploration of the proposed theory of sets and classes, and it is not clear to what extent it will repay further study. Still, we must be grateful to Parsons for the analogy that suggested this possibly illuminating approach.

NOTES

I am indebted to Anil Gupta, John Myhill, and Bill Tait for their suggestions and observations concerning my 1983 article, and to Gupta and Tait again for numerous helpful suggestions on earlier drafts of the present paper (as will become abundantly clear in footnotes to follow). My thanks also to the National Science Foundation, whose grant #SBR-9320220 supported this research.

1. See, for example, Parsons (1974a,b, 1977).
2. Kripke (1975). See also Martin and Woodruff (1975).
3. This is an extension and improvement of my earlier theory (Maddy 1983) (which also contains more detail on the history of treatments of proper classes).
4. See Feferman (1984) for discussion of these earlier theories.
5. I use 'collection' as a term neutral between sets and classes.
6. The theory in Maddy (1983) did not permit free variables within class terms, and so, it was impossible to quantify into class contexts (as in the third example).
7. Though $\overline{V}$ may be redundant, the question of whether or not there is a $t \in T^*$, not involving $\overline{V}$, with $t^+ = \{\bar{a} \mid a \in V\}$ and $t^- = T^* - \{\bar{a} \mid a \in V\}$ remains open. (This addition to the earlier theory was suggested by both Gupta and Myhill.)
8. See Feferman (1984, p. 88). This improvement of the earlier theory was suggested by Myhill.
9. This simultaneous definition, suggested by Gupta, replaces the cumbersome definitions of Maddy (1983).
10. Obviously, I am speaking here of collections too big to be sets; in other words, my account of sets and classes presupposes some rudimentary understanding of classes, much as the iterative picture of sets presupposes some rudimentary notion of set.
11. The symbols '=' and '$\in$' (and others defined in terms of these) are ambiguous here: they appear both in the formulas of the language $\mathscr{L}$, and in meta-linguistic uses (as in this paragraph). The context will disambiguate.
12. The earlier theory lacked '=' as a primitive, because I had not realized that this trivial definition would work. For the shortcomings of less trivial definitions, see Section II.
13. Alas, '$\supset$' behaves rather strangely: $\mathfrak{C} \models p \supset q$ iff $\mathfrak{C} \neq p$ or $\mathfrak{C} \models q$, so if $\mathfrak{C} \overset{?}{\models} p$, then $\mathfrak{C} \models q$. As a result, if $\mathfrak{C} \models p \equiv q$, then $\mathfrak{C}$ cannot be undecided about either p or q. This fact will come back to haunt.
14. For example, to determine whether or not $\mathfrak{C} \models x \in \hat{y}(y \in z)[s]$, where $s(x)$ is $\emptyset$ and $s(z)$ is $(\hat{x}(x \in y), y \to \{\emptyset\})$, we must first remove the clash of variables by replacing $(\hat{x}(x \in y), y \to \{\emptyset\})$ with $(\hat{u}(u \in v), v \to \{\emptyset\})$; second, substitute $\hat{u}(u \in v)$ for z; third, adjust s to s' where $s'(x)$ is $\emptyset$ and $s'(v)$ is $\{\emptyset\}$; and fourth (to see whether or not $\mathfrak{C} \models (x \in \hat{y}(y \in \hat{u}(u \in v))[s'])$, determine whether or not $\emptyset$ is in $(\hat{y}(y \in \hat{u}(u \in v)), v \to \emptyset)^{\dagger}_{\mathfrak{C}}$.
15. Any of the usual set-theoretic definitions of ordered pairs can be imitated here, but Gupta, again, pointed out that this extremely simple idea also works.
16. Tait bemoaned the absence of such class relations in my (1983) and suggested that they be added outright, but the presence of the trivial '=' relation makes this unnecessary. Tait's fixed-point theorem (in Section III, below) depends on class relations; the validity of the theorem for the earlier theory remains open.
17. The corresponding proof in Maddy (1983) could be carried over, but the added flexibility of the extended system makes a more direct proof possible.
18. 'A' domain or range because of the failures of extensionality discussed in Section II.
19. If we identified the last two, we would need to complicate the definition of (x, y).
20. Cf. Kripke (1975, p. 704).
21. The oddities of $\supset$ come into play here, but they will not affect what follows, because we concentrate on ordinals, which are total.
22. Tait conjectured and Flagg later proved that if the construction is carried out over a ground model of ZF with real ordinals, then it will reach a fixed point at the first admissible ordinal greater that the least upper bound of the ordinals in the ground model.

23. And thus, '$t \approx t$' is another way of expressing the claim that t is total, that is, $U \models t \approx t$ iff $U \models \forall x(x \in t \vee x \notin t)$.
24. 'Fregean' because numbers are identified with equivalence classes; 'quasi-Fregean' because of the element of arbitrariness that enters with the conspicuous use of the von Neumann ordinals. This last would be less bothersome if, for example, $\hat{x}(x \approx \overline{\{\emptyset, \{\emptyset\}, \{\emptyset, \{\emptyset\}\}\}}) = \hat{x}(x \approx \overline{\{\emptyset, \{\emptyset\}, \{\{\emptyset\}\}\}})$, but our definition of '=' is too fine-grained for this.
25. See Maddy (1990, Ch. 3), for a philosophical discussion of this approach. The particular approach taken here is suggested at the end of Maddy (1983).
26. See Goldstein (1983).

REFERENCES

Feferman, S. (1984). "Toward Useful Type-Free Theories, I,"*Journal of Symbolic Logic*, 49: 75–111.

Flagg, R., and Myhill, J. (1987). "Implication and Analysis in Classical Frege Structures," *Annals of Pure and Applied Mathematics*, 34: 33–85.

Goldstein, R. (1983). *The Mind-Body Problem* (New York: Random House).

Kripke, S. (1975). "Outline of a Theory of Truth," *Journal of Philosophy*, 72: 690–716.

Maddy, P. (1983). "Proper classes," *Journal of Symbolic Logic*, 48: 113–39.

Maddy, P. (1990). *Realism in Mathematics* (Oxford: Oxford University Press).

Martin, R., and Woodruff, P. (1975). "On Representing 'True in L' in L," *Philosophia*, 5: 113–17.

Parsons, C. (1974a). "Sets and Classes," reprinted in Parsons (1983), pp. 209–20.

Parsons, C. (1974b). "The Liar Paradox," reprinted in Parsons (1983), pp. 221–67.

Parsons, C. (1977). "What Is the Iterative Conception of Set?" reprinted in Parsons (1983), pp. 268–97.

Parsons, C. (1983). *Mathematics in Philosophy* (Ithaca, NY: Cornell University Press).

Challenges to Predicative Foundations of Arithmetic

SOLOMON FEFERMAN* AND GEOFFREY HELLMAN

> The White Rabbit put on his spectacles. "Where shall I begin, please your Majesty?" he asked. "Begin at the beginning," the King said gravely, "and go on till you come to the end: then stop."
>
> Lewis Carroll, *Alice in Wonderland*

This is a sequel to our article "Predicative foundations of arithmetic" (Feferman and Hellman, 1995), referred to in the following as PFA; here we review and clarify what was accomplished in PFA, present some improvements and extensions, and respond to several challenges. The classic challenge to a program of the sort exemplified by PFA was issued by Charles Parsons in a 1983 paper, subsequently revised and expanded as Parsons (1992). Another critique is due to Daniel Isaacson (1987). Most recently, Alexander George and Daniel Velleman (1996) have examined PFA closely in the context of a general discussion of different philosophical approaches to the foundations of arithmetic.

The plan of the present paper is as follows: Section I reviews the notions and results of PFA, in a bit less formal terms than there and without the supporting proofs, and presents an improvement communicated to us by Peter Aczel. Then, Section II elaborates on the structuralist perspective that guided PFA. It is in Section III that we take up the challenge of Parsons. Finally, Section IV deals with the challenges of George and Velleman, and thereby, that of Isaacson as well. The paper concludes with an Appendix by Geoffrey Hellman, which verifies the predicativity, in the sense of PFA, of a suggestion credited to Michael Dummett for another definition of the natural number concept.

I. Review

In essence, what PFA accomplished was to provide a formal context based on the notions of finite set and predicative class and on prima facie evident principles for such, in which could be established the existence and categoricity of

*This paper was written while the first author was a Fellow at the Center for Advanced Study in the Behavioral Sciences (Stanford, CA) whose facilities and support, under grants from the Andrew W. Mellon Foundation and the National Science Foundation, have been greatly appreciated.