8

Circularity and Paradox

Stephen Yablo

1

Both the paradoxes Ramsey called semantic and the ones he called set-theoretic look to be paradoxes of *circularity*. What does it mean to say this? I suppose it means that they look to turn essentially on circular notions of the relevant disciplines: semantics and set theory. But what does it mean to call a notion circular? I suppose that a *circular* notion (of discipline D) is one of the form *self-*R, for R a key relation of that discipline.

Reference and predication are key semantic relations, so self-reference and self-predication are circular notions of semantics. Membership is a key set-theoretic relation, so self-membership is a circular notion of set theory. The set and semantic paradoxes look to be paradoxes of circularity because they look to turn essentially on notions like self-membership and self-reference.

This approach to circularity might seem insufficiently discriminating. Do we really want to count self-deception and self-incrimination in with self-reference and self-membership?

Well, why not? Remember, the target here is not circular notions as such but circularity-based paradox. We get a circularity-based paradox when a circular notion generates absurdities, with the circularity of the notion playing an essential role. I don't know whether self-deception and self-incrimination generate absurdities in this way. But if they do, then I for one am happy to speak of circularity-based paradoxes of psychoanalysis or legal theory.

Self-Reference.
Thomas Bolander, Vincent F. Hendricks,
and Stig Andur Pedersen (eds.).

No one imagines that circularity generates paradox all by itself. The claim is that circularity is *necessary* for the relevant sorts of paradox to arise, not that it is *sufficient*. A paradox is circularity-based if circular notions wreak havoc when deployed in a particularly nasty way, and corresponding non-circular notions cannot be made to wreak similar havoc.

2

Our main question in this paper is: Are the semantic and set-theoretic paradoxes circularity-based? This has been for a long time the dominant view. It shows up in the frequently heard claims that one sure way to avoid the semantic paradoxes is to insist with Tarski on a rigid separation of object language from meta-language, and one sure way to avoid the set paradoxes is to insist with Russell on a rigid hierarchy of types.

But these claims are open to question, especially the first. Tarskian strictures may block the Liar paradox but they do not block all paradoxes of the Liar type. An example is what we can call the ω-Liar. This involves an infinite series of sentences S_i, each describing as false all S_js occurring later in the sequence:

$$\begin{aligned} S_0 &= \forall n \geq 1 \sim T[S_n] \\ S_1 &= \forall n \geq 2 \sim T[S_n] \\ &\vdots \\ S_i &= \forall n \geq i+1 \sim T[S_n] \\ &\vdots \end{aligned}$$

Earlier S_is entail later ones, so if any S_i is true so are all the ones after it. At the same time S_i is true only if the sentences after it are *false*. Therefore no S_i can be true; the only consistent assignment makes them all false. However that assignment is not consistent either, since now the truth conditions of each S_i are fulfilled. So we have an intuitive contradiction.[1]

How the ω-Liar can arise in a Tarskian setting may not be immediately obvious. The answer lies in an observation Kripke makes in 'Outline of a Theory of Truth':

[1] Whether you can *prove* the contradiction depends on the logical resources available, but that's another matter. All we need for paradox is that the S_is cannot be consistently evaluated on their intended interpretation.

> One surprise to me was the fact that the orthodox approach by no means obviously guarantees groundedness ... standard theorems easily allow us to construct a *descending* chain of first order languages $L_0, L_1, L_2, \ldots$ such that L_i contains a truth predicate for L_{i+1}. I don't know whether such a chain can engender ungrounded sentences, or even quite how to state the problem here; some substantial technical questions in this area are yet to be solved (Martin (1984), 61).

Suppose that each L_i contains a sentence S_i of the type indicated above, each describing its successors in lower-level languages as untrue. S_0 contains a truth predicate T_0 that captures the truths of language L_1, L_2, etc.[2] S_1 contains a truth predicate T_1 that captures the truths of $L_2, L_3, \ldots$ And so on.[3]

Now the argument goes much as before. If S_5, for instance, is true_4, then S_6, S_7, and so on are false_4. But then S_6 ought instead to be true_4, contradiction. Therefore no S_{i+1} is true_i. But then each *should* be true_i, since its successors are false_i. So the Tarskian way of avoiding paradox relies on more than a rigid object/metalanguage distinction. It is also required that the sequence of languages eventually grounds out in a bottom-level object language.

I call the ω-Liar a non-circular paradox, but one might question this. Where is the proof that there really are sentences of the kind described? Gödel went to a lot of trouble to establish the existence of even a single self-referential sentence, yet I am just assuming a whole infinite string of sentences each referring to all the ones after.

Gödel goes to all of this trouble for a reason, however. He wants to establish a result about existing arithmetical theories (that they are inconsistent or incomplete). He knows how to do this for theories with certain expressive powers: theories that can express (1) the concept of provability, and (2) the thought, for each expressible concept C, that *this very sentence does not fall under* C. This gives Gödel the result he wants only if Peano Arithmetic (e.g.) has the expressive powers in question. Naturally then he works hard to show that it does.

Compare the situation faced by someone like Tarski.[4] He too seeks to establish a result about consistent theories with certain expressive powers. But the result is not that theories of the relevant type are

[2]This may seem to depart from Kripke's picture, since for him the truth predicate of L_i applies only to sentences of L_{i+1}. But if metalanguages extend their object languages, then the sentences of L_{i+1} will include those of L_{i+j} for all $j \geq 1$.

[3]I owe this point to Forster (unpublished) and Visser (1989).

[4]'Like' because I do not say that this is how Tarski himself judges the situation.

incomplete; it's that there *are* no such theories. No consistent theory with power (2) can express (1′) its own concept of truth. Gödelian subtleties are unneeded here because it doesn't matter for Tarski's purposes whether self-reference is accomplished through arithmetization of syntax or by some other method. Similarly it doesn't matter for our purposes how we arrange for the ω-Liar's patterns of other-reference.

One approach is to suppose that the language has predicates $P_1, P_2, \ldots$ whose extensions $E_1, E_2, \ldots$ are stipulated. Among the sentences are $\forall x(P_1x \rightarrow \sim Tx), \forall x(P_2x \rightarrow \sim Tx), \forall x(P_3x \rightarrow \sim Tx), \ldots$ What is to prevent us from stipulating that P_1 shall be true of the first of these? Or that P_i applies to the sentences of this form containing P_k for $k > i$?

It might be argued that what prevents it is that the indicated assignments generate paradox. But the idea that reference relations are subject to this kind of sentence-level constraint is highly implausible. Imagine that it is not extensions E_i that are stipulated but extension-determining properties P_i. Then we might discover empirically that P_i has a paradox-making extension; or we might shove sentence-tokens around to make it so. Tarski himself emphasizes the form of the Liar that turns on the 'empirical premise' that the sentence written on the board in room 301 is 'The sentence on the board in room 301 is false.' How an empirical description could fail to apply because of an object's semantical properties it is not easy to see.

The ω-Liar can also arise for empirical reasons. The sentence on the board in room 301 might be 'the sentences on the boards in rooms 302, 303, etc. are false,' with analogous sentences on the boards in those other rooms. If indexicality is allowed it can be the *same* sentence (type) written on each board, viz. 'the sentence-tokens on the board in those rooms [insert arrow here] are all false.' Roy Sorensen's version has an infinite queue of students, each thinking 'the beliefs of *those* people [pointing back] are all false.'[5]

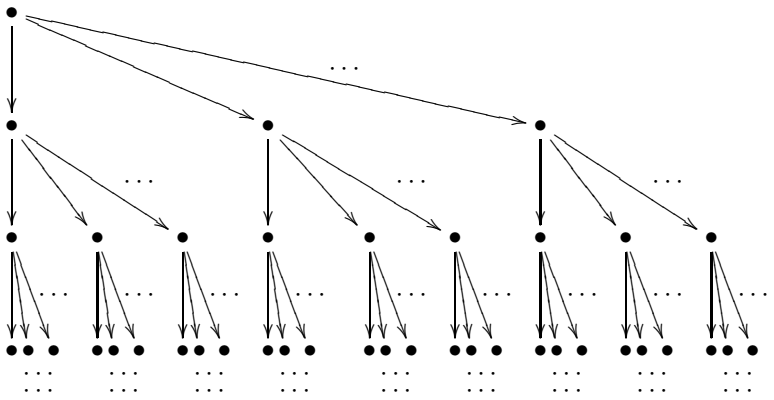
The demonstrative form of the paradox—the beliefs of *those* people are false—gives rise to a different worry. If everyone is in structurally speaking the same situation—each stands at the front of an infinite string of people each thinking 'the beliefs of the people back there are false'—how are the various thoughts to be distinguished? One might even worry that everyone in the line is thinking in some sense *the very same thing*.[6] Thus Priest:

[5]Sorensen (1998).

[6]This *does* seem a funny thing to start worrying about now, when the literature is full of examples of 'indistinguishable' statements that do not collapse into one (for instance, Liar cycles $A_0, \ldots, A_{n-1}$ where each attributes falsity to its modulo

> There would appear to be no circularity here. But there is. This is most obvious if one individuates thoughts in such a way that all the people are thinking the same thought, t. If this is the case, then the thought that they are thinking is just equivalent to the thought that t is not true. The circularity is obvious. In fact, this is just a variant of the liar paradox (Priest (1997), 240).

A point in favor of the structural collapse worry is that if we try to model the propositions involved in Aczel's non-well-founded set theory, they come out identical. This is because Aczel has one set per isomorphism-type of directed graph, and the graphs here are isomorphic; each has the structure of a downward facing tree with omega branches descending from each node.



But even granting Aczel's assumptions, we can arrange for the propositions to be distinct. The graphs are isomorphic because each sentence calls *all* later sentences false. But the paradox doesn't require this. It is enough, for instance, if each S_i says 'all my successors other than [insert here a finite list of exceptions] are false.' Thus we can have the first sentence say 'leaving aside the very next sentence, all my successors are false,' the second say 'ignoring the next sentence but one, all my successors are false,' and so on. This modified sequence is still paradoxical, and now the propositions are distinct even by Aczel's lights.

The conditions can be weakened further. Call S_i cofinite iff it is truth-conditionally equivalent to a (possibly infinite) disjunction of the 'finite list of exceptions' sentences just considered, so that it is sufficient for S_i's truth that every subsequent sentence is false, and necessary for

---

n successor).

its truth that all but finitely many subsequently sentences are false. A sequence of S_is all but finitely many of whose members are cofinite is called cofinite too.

Claim. Cofinite sequences are paradoxical.

Proof. Suppose some cofinite S_n is true. Then all but a finite number of later sentences are false; so for some p, S_{n+p} and everything later are false. Let S_{n+p+k} be the next cofinite sentence after S_{n+p}. S_{n+p+k} is false but it should be true, since everything after it is false. So all cofinite sentences are false, whence all sentences are false once we get past the last non-cofinite sentence S_j. Each S_i $(i > j)$ should be true, for it is followed just by falsehoods. Contradiction.

Paradox need not result if (only) *infinitely* many S_is are cofinite, or indeed if infinitely many S_is call *all* subsequent sentences false. All of these can consistently be evaluated as false, provided the sequence's infinitely many *other* sentences are true (the other sentences might each say that *some* subsequent S_is are false). It also does not suffice for paradox if cofinitely many S_is describe *infinitely* many successors as false. For instance let S_i be

$$\forall k > i(k \text{ is a power of the } i\text{th prime} \rightarrow S_k \text{ is false})$$

if i is not the power of any prime, and

$$\forall k > i(S_k \text{ is false})$$

if i is the power of some prime. Then we can consistently call S_i false when i is the power of a prime and true when i is not a prime power. So the above looks like the strongest result based just on cardinality considerations.

3

So much for the semantic paradoxes, and the idea that Tarskian strictures are enough to keep them at bay. The closest set-theoretic analogue of Tarski's theory is the simple theory of types. This was a response to the discovery that naïve set theory is inconsistent. We can take naïve set theory to be a two-sorted first-order theory with distinct variables for sets and properties. It has two non-logical symbols—$\in$ for membership and 'has' for predication—and two axioms. Extensionality says that sets with the same members are identical:

$$\text{EXT} \qquad \forall x \forall y(\forall z(z \in x \leftrightarrow z \in y) \rightarrow x = y)$$

Naïve comprehension says that for any property P there is a set of things possessing the property:

NCO $\quad \forall$ properties $P\, \exists y \forall z (z \in y \leftrightarrow z \text{ has } P)$

What makes the comprehension naïve is that there is no theory of legitimate or kosher properties hovering in the background; a property is whatever makes intuitive sense as such. Non-self-membership makes intuitive sense, so naïve comprehension tells us there ought to be a set R containing all and only sets that don't belong to themselves. The rest is history.

The response as I understand it has two parts, a negative part and a positive one. The negative response is to complain that in constructing the Russell set R as above one is violating the Vicious Circle Principle:

> Given any set of objects such that, if we suppose the set to have a total, it will contain members which presuppose this total, then such a set cannot have a total. By saying that a set has 'no total,' we mean, primarily, that no significant statement can be made about all its members (Whitehead and Russell (1910), 37).

How does this help? The set of non-self-membered sets, if we suppose it to have a total, must contain itself, since any other hypothesis leads to contradiction. A set that contains itself has a member which presupposes the set's entire membership, because it *is* the set's entire membership gathered into a set. So the Russell set has no total, which we can take to mean that there is no such set.

Russell however wanted to give a *positive* theory that would *automatically* sidestep totalities at variance with the vicious circle principle. He insists that 'The exclusion must result naturally and inevitably from our positive doctrines' (Russell (1956)). He envisages a hierarchy of ever-increasing types, where sets of type n are the extensions of properties of sets of type less than n. Naïve comprehension is restricted to say:

> for any P a property of *sets of a given type*, there is a set one type up containing all and only the Ps of that type.

Formally we can think in terms of an ω-sorted language, where variables of type n are marked with a superscript n. Naïve comprehension now becomes

NCO^{n+1} $\quad \forall P \exists y^{n+1} \forall x^n (x^n \in y^{n+1} \leftrightarrow x^n \text{ has } P)$.

Once again this is not backed by any general theory of properties; our freedom to dream up properties is subject to no other limits than the

one indicated. One could even say that the properties are whatever they were before; it is just that in applying comprehension one forms at the $(n+1)$st level the set of all nth-level objects possessing the property in question. (Extensionality is similarly modified and an axiom of infinity is added to ensure infinitely many objects of the same type.)

Now as type theory is *usually* understood, it has individuals of type 0, classes (really prop. functions, but we simplify) of individuals at the 1st level, classes of classes of individuals at the 2^{nd} level, and so on. What does this mean for the Russell set? There is a level 1 Russell set R^1 containing all level 0 entities that aren't members of themselves, a level 2 Russell set R^2 containing all level 1 entities that aren't members of themselves, and so on. How is the contradiction avoided? Well, that a Russell set R^{n+1} does not belong to itself ceases to mean that oh yes it does, because the property defining membership in R^{n+1} is not non-self-membership but that *plus* being a set of level n. The $(n+1)$st-level Russell set satisfies the first condition but not the second.

So on the one hand we've got the vicious circle principle, and on the other hand, we've got the type structure just sketched. The type structure involves *two* ideas—one, *hierarchy of types*, two, *starting from type* 0—of which only the first can claim any sort of support from the vicious circle principle. If circularity is really at the heart of the matter, the first idea—hierarchy of types—should be enough to make things right. It should be consistent to allow infinitely descending types, where as before the sets at level $(n+1)$ are the extensions of properties of level n sets. Naïve comprehension takes exactly the same form as before except that now we allow n to range over *all* integers, not just the natural numbers.

A system like this does occur in the literature—initially in an article by Hao Wang in *Mind* 1952 called 'Negative Types,' but with a trickle of sightings later on, for instance in an article by Ernest Specker (in the *Proceedings of the 1960 International Congress in Logic, Methodology, and Philosophy of Science*) called 'Typical Ambiguity.' Thomas Forster calls the theory TNT and I will use his terminology.[7]

What is going on in the articles by Wang and Specker? Both authors distinguish the naïve (Specker calls it 'ideal') form of the theory, which has unrestricted comprehension at each level, from the formalized version in which we have instead an axiom schema giving us, for each formula $\varphi(x^n)$ with a free variable of type n, a set of type $(n+1)$ containing all and only the φs. Both observe that *formalized* TNT is consistent if regular type theory is, by a compactness argument. Wang

[7] See Forster (1995).

observes about *naïve* TNT that it has no need of the axiom of infinity, since comprehension already provides images of all lower level null sets; level 0 for instance contains $\{\emptyset_{-1}\}, \{\{\emptyset_{-2}\}\}, \{\{\{\emptyset_{-3}\}\}\}$, and so on.

TNT serves as a test case for Russell's diagnosis of the set paradoxes. If they result from violations of the Vicious Circle Principle, then since naïve TNT respects the principle, paradoxes should not arise in it. Naïve TNT does appear safe from Russell's paradox. But a counterpart of Mirimanoff's grounding paradox arises.

Mirimanoff's paradox (for naïve set theory) is this: A set is well-founded if it heads no infinite descending epsilon chains; it is not the case that x has a member which has a member which has a member and so on indefinitely. Some sets are well-founded, and others aren't; the universal set U, for instance, is not well-founded since $U \ni U \ni U \ni U \ni \cdots$.

Consider the set W of all well-founded sets. On the one hand it is well-founded, because an infinite descending chain from W requires an infinite descending chain from one of its members, and that is impossible as its members are well-founded. But if as this argument shows W is well-founded, then it belongs to the set of all well-founded sets, that is, it belongs to itself, which makes it not well-founded after all.

The original Mirimanoff's paradox is a paradox of circularity. But the variant that arises in naïve TNT is not circular. NCO^n has for each integer n a set W^n of well-founded sets of level $n-1$. Each W^n is well-founded for the same reason as before, namely that if not then one of its members heads an infinite descending $\in$-chain, contradicting the fact that its members are well founded. Since W^n is well-founded, each W^n belongs to the set of well-founded sets of level n, that is, it belongs to W^{n+1}. Now we have shown that for all n, $W^n \in W^{n+1}$. Letting n run down through the negative integers from 0 to -1 to -2 etc., we get that $W^{-1} \in W^0, W^{-2} \in W^{-1}, W^{-3} \in W^{-2}$, and so on ad infinitum. Putting the pieces together, $W^0 \ni W^{-1} \ni W^{-2} \ni W^{-3} \ni \cdots$. But then W^0 isn't well-founded after all. Contradiction.

A skeptic might say that well-foundedness has shown itself not to be a well-defined property in the context of (naïve) TNT, since when we try to work out which objects possess it we get tangled in knots. But the same could be said about the property of non-self-membership in the context of naïve set theory. The Vicious Circle Principle was supposed to protect us from this sort of tangle, and yet here we have sets constructed in conformity with the principle and the tangle remains. Rigid type separation is helpful, but it cannot restore consistency all by itself. One needs to assume in addition that the hierarchy of types bottoms out.

4

I believe that the ω-Liar is fairly convincing against the view that semantic paradox requires circularity. Our modified Mirimanoff is much less convincing against the corresponding view about set paradoxes. It is not hard to see the reason for this. The ω-Liar is a noncircular paradox for *intuitive* theories of truth, not just the cooked-up theory you get by foisting infinite descending object languages on Tarski. And although Mirimanoff does arise for intuitive theories of sets, it is *noncircular* only in connection with the cooked-up theory TNT. Do noncircular paradoxes arise for any *attractive* theory of sets?

An attractive set theory requires the sets to form a well-founded structure. The universe of the theory should be a cumulative hierarchy, formed by starting with urelements and then repeatedly gathering together the things already on board into sets. It is true that we cannot run the modified Mirimanoff paradox in this setting. But perhaps we don't have to. A case can be made that Russell's paradox is not itself, in this setting, a paradox of circularity.

A set-theoretic paradox arises when a set that, naively speaking, ought to exist cannot exist, because the hypothesis of its existence conflicts with fundamental facts about sets. An explanation of the paradox is an explanation of how and why the conflict arises. Such an explanation might take either of two forms.

One sort of explanation shows us why the *de dicto* hypothesis that there is a set of Fs conflicts with basic facts about sets. The reason for this might be different from the reason there is no set of Gs, even if it is agreed all around that every F is G and vice versa. It is provable in ZF that $x \notin x$ iff $\forall y \sim(x \in y \& y \in x)$ iff $x = x$. But that doesn't mean that we explain the non-existence of $\{x \mid x \notin x\}$, $\{x \mid \forall y \sim(x \in y \,\&\, y \in x)\}$ and $\{x \mid x = x\}$ the same way.

A second sort of explanation tries to show why the quasi-*de re* hypothesis of a set with such and such a nature cannot be reconciled with basic facts about sets. What is it about the putative set considered in itself that most fundamentally prevents it from existing?

I admit that it is not always obvious which sort of explanation we are after, when we ask why a certain paradox arises. And sometimes a *de dicto* explanation is the most we can hope for, because the set's defining condition affords few clues to its membership. But it seems to me that where the set's putative membership is clear, the *de re* explanation is more revealing. One wants to know not why a set thus and so specified cannot exist, but why a set of such and such a nature—with such and such members—cannot exist.

Consider in this light some explanations that might be given of what prevents the Russell set $\{x \mid x \notin x\}$ from existing. Russell himself tells us that

(I) $R \in R \longrightarrow R$ fails to meet R's membership condition $\longrightarrow R \notin R$.
$R \notin R \longrightarrow R$ does meet R's membership condition $\longrightarrow R \in R$.

This is fine if (like Russell) we have no clear idea of the pool of sets from which R's membership is to be drawn. But from our present perspective, it seems objectionably de dicto.

The reason is this. The feature of R that (I) finds trouble with—containing all sets that fail to contain themselves—is of no intrinsic importance; the reasons Russell had for stressing this feature no longer apply, now that we are conceiving sets as well-founded. Meanwhile the fact of what R as an extensional entity is, or would be, is ignored. One might as well say that what prevents the integer 998/3 from existing is that 26, the sum of its decimal-representation digits, is not divisible by 3.[8]

The lesson we draw from (I) is that our notation for a thing cannot be the reason it does or does not exist. Can an explanation be found that does not play quite so dumb about the fact that R would have to be $U =$ the set of all sets?

(II) U is a set $\longrightarrow U$ meets U's membership condition $\longrightarrow U \in U$.
U is a set $\longrightarrow U$ is well-founded $\longrightarrow U \notin U$.

This brings what the Russell set is by nature into conflict with a basic fact about sets, viz. well-foundedness. But one may question whether the fact is basic enough. To say that U cannot exist because it would be ill-founded seems to get things the wrong way around. It is because sets like U are *independently* problematic that we are drawn to a requirement that keeps those sets out.

What is the deep and underlying problem with U, if not that it would have to belong to itself? The problem is that U is an example of a *type* of set *all* of whose instances should belong to U. I refer, of course, to U's subsets. U being universal ought to contain all of these. But it can't, because a set has more subsets than members. Here we reach a constraint that cannot be qualified or tinkered with, because it is a second order logical truth that

$$\sim\exists R(R \text{ maps the pluralities of objects 1-1 into the objects}).$$

This takes some encoding, to be sure. R ranges over relations on individuals, so what can it mean to say it takes a *plurality* of objects—the

[8] Relying here on the fact that n is divisible by 3 iff 3 divides the sum of the digits in its decimal.

Fs—to some *particular* object o? Let $R(F,o)$ be short for $\forall u(Ruo$ iff u is one of the Fs). The logical truth is

$$\sim\exists R\forall F\exists o\big(R(F,o)\ \&\ \forall x(R(F,x)\rightarrow x=o)\ \&\ \forall G(R(G,o)\rightarrow F=G)\big),$$

unpacked in the manner indicated. This is the logical core of Cantor's Theorem, and it gives what I take to be the real problem with a universal set: it would have to contain a distinct object for each plurality of objects (viz. the set of objects), and that is logically impossible.

(III) U exists $\longrightarrow$ all sets form a set $\longrightarrow$ any sets form a set.[9] U contains all sets $\longrightarrow$ U contains all of U's subsets $\longrightarrow$ $\sim$(U's subsets outnumber its members) $\longrightarrow$ U has members that do not form a set $\longrightarrow$ $\sim$(any sets form a set).

For factor X to be crucial to set theoretic paradox, it is not enough that X should wreak havoc; it should wreak havoc *not because of being associated with other factors that would wreak havoc all by themselves.* What does this say about Russell's paradox, considered as a paradox of circularity? The Russell set certainly wreaks havoc. But then R is the universal set, and for a set to contain everything wreaks havoc all by itself. This is what makes me at least suspicious of the idea that Russell's paradox is a paradox of circularity.

5

Since presumably the universal set *would* exist if Naïve Comprehension held, Naïve Comprehension looks to be the source of our problems. I believe, however, that Naïve Comprehension (suitably interpreted) *does* hold, and can serve once again as the main engine of set production.

One normally thinks of NCO as tied into the 'logical' conception of a set: sets as extensions of predicates, or properties. The failure of NCO in that context leads us to replace the logical conception with the 'combinatorial' conception, which says that sets are formed by gathering together other sets that were earlier formed in the same way. A version of comprehension remains—separation—but it is just one more axiom (or schema). It is no longer what drives the whole process ahead as on the naïve theory.

[9] Unless some of U's subsets are 'missing,' in that although *all* of U's members form a set, there are *some* of its members that fail to do so. This conflicts however with a *very* fundamental feature of sets, namely (downward) closure: if there is a set of all Ys, and the Xs are some of the Ys, then there should be a set of all Xs. Well-foundedness is valued basically for hygienic reasons, but closure is part of what we have in mind by a set.

I submit that dropping the logical conception of set does not have to mean dropping Naïve Comprehension. Moreover the role that Naïve Comprehension tried to play on the logical conception, it can succeed in playing on the combinatorial conception. NCO can in some sense resume its rightful place as the principal engine of set production and pretty much the sole axiom of set theory apart from Extensionality.

How does this go? Naïve comprehension is usually treated as a principle about properties (or concepts) and sets. There is room for a third player between these two, namely *pluralities*. When pluralities are crowbarred in, NC appears as the product of *two* principles.

> *Naïve Plurality Comprehension* (NPC)
> For any property P, there are the things that are P.
>
> *Naïve Set Comprehension* (NSC)
> Whenever there are some things, there is the set of those things.

These are not very often separated but when they are, the blame is generally put on Naïve Set Comprehension. This is the 'limitation of size' diagnosis: *some things are too many to form a set*. I propose to put the blame rather on Naïve Plurality Comprehension.

The objection will come that NSC entails the existence of a universal set. Consider *the self-identical things*, or the things with any other trivial property. NSC apparently guarantees that there is a set of these of things, and that will be the set of everything.

This reasoning goes wrong at the first step. There is certainly a *property* of being self-identical, or at least I am not quarreling with that. But the objection further assumes that there are *the things possessing this property*. The point of rejecting Naïve Plurality Comprehension is that we reserve the right to say, yes, there is this property, but no, there are not the things possessing the property. There are, of course, *individual* things possessing it. The first-order statement

> there is a thing x such that x has P

is quite true. What we deny, or reserve the right to deny, is the second-order statement

> there are things the Xs such that $\forall y$(y is one of them iff y has P).

In particular it is not the case that there are some things comprising all and only the self-identical things. The view once again is that *plurality comprehension* is mistaken.

This may seem at first puzzling. The property P that (I say) fails to

define a plurality can be a perfectly determinate one; for any object x, it is a determinate matter whether x has P or lacks it. How then can it fail to be a determinate matter what are *all* the things that have P? I see only one answer to this. Determinacy of the Ps follows from

(i) determinacy of P in connection with particular candidates,

(ii) determinacy of the pool of candidates.

If the difficulty is not with (i), it must be with (ii). It is not the case that there are some things the Xs such that every candidate for being P is among them. If there were, one could go through the Xs one by one, asking of each whether it has P, thus arriving finally at the sought-after plurality of Ps.

How could there fail to be a determinate pool of candidates for being P? The answer lies in the combinatorial conception of sets. The universe of sets is, according to that conception, built up bit by bit, recursively as it were; a set's members come in before the set itself. Kit Fine has a nice device for making this vivid. He asks us to imagine that we have a genie at our disposal. We give the genie instructions and s/he carries them out. Instructions can be simple—take the successor of this number—or iterative—keep on taking successors until further applications produce nothing new. The advice Kit favors in set theory is: take power sets at successor stages and unions at limit stages, until further application yields no new accessible cardinals.

I would like to change this advice in two ways. The first change is that I want my genie to be as dumb as possible, so that the sets are built up by repeated application of the simplest possible instructions. Power sets and unions at limit ordinals are more than my genie can handle. They are replaced by a single instruction:

(∗) whenever you have made some things, make the set of them.

The second change is this. Fine has the genie stopping when all sets of a certain size have been reached. This detracts from overall simplicity because the notion of an accessible cardinal is rather sophisticated. And it adds arbitrariness because the genie could equally well have stopped elsewhere. As Hellman says,

> there are ways of restricting the heights of models ... [so as to determine set theoretic truth uniquely] ... However there are at least two substantial obstacles to this course. The first would be the arbitrariness of the [Axiom of] Restriction. If the aim is to produce a categorical theory, one can achieve this in infinitely many different ways (e.g. add to ZF^2 the axiom that there are exactly 17 in accessibles, etc.) and no

> evident reason to single out one as optimal (Hellman (1989), 74).

The one principled reason I can think of for stopping precisely *here* would be that you have now got too many objects to form a set. But I agree with Hellman (and Putnam) in finding it incomprehensible how that could happen. What could the obstacle possibly be to gathering the sets created so far into a new set? It conflicts with 'something too deeply rooted in our use of set-like operations to renounce the possibility of "going beyond" any definite totality' (ibid., 74).

This suggests an instruction more like what Fine gives for the natural numbers: keep on going until further applications produce nothing new. The difference is that this time further applications always *do* produce something new. The set of Xs is always additional to the Xs. So in the set case, *the genie's work is never done*. This is not because set-creation is so intrinsically time consuming, but because whatever you might propose as the stopping point affords the genie materials for adding something new. The instruction more fully stated is

($*$) whenever you have made some things, form their set, *continuing forever*.

Note that 'continuing forever' cannot just mean, 'don't ever stop making new sets.' That instruction a lazy genie could follow by making the empty set at $t = 0$, its singleton at $t = 1$, and so on through the rest of time. This would produce only the (Zermelo) naturals. 'Continuing forever' means, and you can consider this stipulative, 'anything a faster-moving genie *could* make, you eventually *do* make.' The lazy genie is not continuing forever in this sense, because a faster-moving genie could make the set of Zermelo naturals, and the lazy genie never does.

6

One would like to be able to show that the genie's evolving universe satisfies the axioms of second-order ZF, with 1^{st} order universal quantifiers interpreted to mean 'for any x that is eventually made...', and 2^{nd} order universal quantifiers interpreted to mean 'for any Fs that are eventually made.' The most I can do here is to pick off some low-hanging fruit.

> *Pairs.* $\forall x \forall y \exists z \forall u (u \in z \leftrightarrow u = x \vee u = y)$.
> Proof: If x and y are eventually made then the genie is instructed to form their set.

Union. $\forall x \exists y \forall z (z \in y \leftrightarrow \exists u \in x\ (z \in u))$.
Proof: If x is eventually made then its members' members must have been made earlier. But then the genie is instructed to form the set of its members' members, and that is $\bigcup x$.

Separation. $\forall F \forall x \exists z \forall y (y \in z \leftrightarrow y \in x\ \&\ Fx)$.
Proof: If x is eventually made, then its members were made earlier. Hence in particular the members of x that are F were made earlier. But then the genie is instructed to form the set of these things.

The null set raises special problems. I see two main options. The first is to say that there are some things the Xs such that nothing is one of them (the things which are non-self-identical), and the null set is the set of these Xs. The second option is to have a special one-time instruction: create a set such that nothing is in it! I will not try to adjudicate between these options today.

Suppose our genie behaves as instructed and no opportunity for set-creation goes untaken. Then no matter what the Xs may be, they do not include every set that eventually gets made. No sets are *all* the sets, because we know of a set not among them, viz. the set with them as its members.[10]

A similar argument shows that no matter what the Xs may be, it is eventually the case that all of them have been created ('all' gets narrow scope). For suppose to the contrary that it is always the case that some X remains to be created. Then the set of Xs is never created. This is a breach of the genie's instructions, since a faster-working genie could have finished off the Xs and then proceeded to make their set. It follows that

(#) things each of which eventually gets made are things that will have eventually all been made.

This brings us back to the promised comprehension principle. The principle I would defend has already been stated

Naïve Set Comprehension
$\forall X \exists y \forall z (z \in y$ iff z is an $X)$.

[10]This is not to say that set theoretic truth is indeterminate. Building on an idea of Zermelo's, Hellman shows how the sequence of ever-larger natural models of second-order ZF can be used to assign truth-values. (A *natural model* is the cumulative hierarchy up to rank α, α an inaccessible. Given any two natural models M and M', one end-extends the other.) Putnam semantics, as this is known, 'gives unique answers to all (ZF) set theoretic questions' (Hellman (1989), 77).

This says that things each of which is eventually made are things whose set is eventually made too. Such a claim would be contradictory if X could assume the value *everything that ever gets made*. But as we have seen, *there are no such things as that*. (#) tells us that there are such things as the Xs only if eventually every X has been made. Thus we could equally put NSC by saying that if eventually every X has been made, then eventually the set of Xs is made. Clearly there is no contradiction in that.

7

How does this bear on anything real? The genie story is only a story. It is not as if the universe of sets is essentially always under construction in the manner envisaged. It is not even clear that a genie 'no more productive than which can be conceived' *could* exist.

But we know from Kreisel that 'real' is said in two ways: there is the reality of entities, and the reality (objectivity) of truth. Kreisel suggests that a reality of the second kind need not always to be backed by a reality of the first kind, and I suspect many philosophers would agree that there is no absolute requirement here.

But it might still be thought desirable, other things equal, for real mathematical truth to have some sort of platonic backing; arithmetical statements, for instance, might seem more objectively right (or wrong) if answerable to an independently constituted domain of natural numbers. I have my doubts about this as a reason for finding the real existence of mathematical objects desirable,[11] but my point here is different. Set theory is the rare case where platonic backing is positively *undesirable* if objective truth is the goal. A statement cannot be objectively true if its truth is arbitrary, and the truth about a chopped-off hierarchy is bound to be arbitrary, for there is no conceivable reason why *these* things should be the first not gatherable into a set.

8

This paper has taken a firmer line on the semantic paradoxes than the set paradoxes. About the first we said that naïve truth theory faces a paradox very much like the Liar that make no essential use of circular notions. About the second we said more or less the following.

(1) A certain oddball set theory invented by Wang faces a paradox sort of like Mirimanoff's that makes no essential use of circular notions.

[11] Yablo (2000).

(2) Russell's paradox as it arises for naïve well-founded set theory is not really a paradox of circularity either.

(3) No known paradoxes beset naïve well-founded set theory, conceived as a theory not about real sets but whatever results from a certain fictitious process.

(4) Conceiving sets the second way lets us hold onto the two most cherished elements in our intuitive thinking about them: a set of Fs for each bunch of Fs ($\forall F \exists y \forall z(z \in y \leftrightarrow Fz)$) and a non-arbitrary truth-value for each set-theoretic hypothesis.

References

Beall, J. C. 2001. Is Yablo's paradox non-circular? *Analysis* 61(3):176–187.

Bringsjord, S. and B. Van Heuveln. 2003. The 'mental eye' defense of an infinitized version of Yablo's paradox. *Analysis* 63(1):61–70.

Forster, T. (unpublished). Yablo's paradox without self-reference. `http://www.dpmms.cam.ac.uk/~tf/yablondjfl.ps`

Forster, T. 1995. *Set Theory with a Universal Set*, 2. ed. Oxford: Oxford University Press.

Goldstein, L. 1994. A Yabloesque paradox in set theory. *Analysis* 54(4):223–227.

Hardy, J. 1995. Is Yablo's paradox Liar-like? *Analysis* 55(3):197–198.

Hellman, G. 1989. *Mathematics without Numbers: Towards a Modal-Structural Interpretation*. Oxford: Oxford University Press.

Kripke, S. 1975. Outline of a theory of truth. *Journal of Philosophy* 72:690–716, reprinted in Martin (1984).

Martin, R. L. 1984. *Recent Essays on Truth and the Liar Paradox*. Oxford: Oxford University Press

Priest, G. 1997. Yablo's paradox. *Analysis* 57(4):236–242.

Putnam, H. 1967. Mathematics without foundations. *Journal of Philosophy* 64:5–22

Russell, B. 1956. Mathematical logic as based on the theory of types. In *Logic and Knowledge; essays, 1901-1950*. London: G. Allen & Unwin.

Specker, E. 1962. Typical ambiguity. In E. Nagel, P. Suppes, and A. Tarski, eds., *Logic, Methodology, and Philosophy of Science: Proceedings of the 1960 International Congress*. Palo Alto: Stanford University Press.

Sorensen, R. A. 1998. Yablo's paradox and kindred infinite Liars. *Mind* 107:137–155.

Tennant, N. 1995. On paradox without self-reference. *Analysis* 55(3):199–207.

Visser, A. 1989. Semantics and the Liar paradox. In Gabbay, D. M. and F. Guenthner (1989), *Topics in the Philosophy of Language*. Dordrecht: D. Reidel.

Wang, H. A. O. 1952. Negative types. *Mind* 61:366–8.

Whitehead, A. N. and B. Russell. 1910. *Principia Mathematica (Vol. I)*. Cambridge: Cambridge University Press.

Yablo, S. 1993. Paradox without self-reference. *Analysis* 53(4):251–252.

Yablo, S. 2000. Apriority and existence. In P. Boghossian and C. Peacocke, eds., *New Essays on the A Priori*. Oxford: Oxford University Press