DEFINITIONS -- CONSISTENT AND INCONSISTENT*

S. Yablo
U. of Michigan, Ann Arbor

published in Philosophical Studies 1993

I. Introduction

"Defining a word is not asserting anything but stating a rule or policy for the word's employment." Some such view of definitions is widely accepted, but what is the philosophical payoff is if it is right? This is where I have a suggestion to make. Show me an inconsistent assertion, and I will show you a false assertion. But an inconsistent rule is not false; indeed it may be correct in the only sense that matters, that of according with speakers' semantic intentions.[1] This opens up the possibility of definitions that are no less correct for being contradictory. Much later I'll suggest that this possibility obtains for a certain definition of truth. Meanwhile we should ask: what is it for a definition qua rule of usage to be inconsistent?

Rules of usage are commonly conceived as putting conditions on objects, the conditions an object must meet for for the word to apply to it.[2] But it seems equally true that they make demands on subjects, telling those who use the word what sort of conduct with it is or is not permitted. This has an interesting and underappreciated consequence, namely that "inconsistent," said of a definition, can mean two different things. The more usual meaning is that the demands the definition places on objects are logically unsatisfiable, as when a glub is defined as a round square. The other, and neglected, alternative is for the definition to impose irreconcilable obligations on speakers. What we lack is an account of definitions that makes room for the second sort of inconsistency. Such an account lies waiting, I claim, just beyond the existing theory of inductive definitions.

II. The Problem

Imagine that we face the task of introducing a novel predicate P into a language L we already understand. How should this be done? Most obviously by providing P with a definition: a rule coordinating P with some combination of L's other vocabulary. Assuming for convenience that L's syntax is first-order, or first-order-like, this rule can be written

$$\Delta \qquad P\underline{x} =_{df} \phi(\underline{x}),$$

where $\phi(\underline{x})$ is a formula with variable $\underline{x}$ free and containing no other free variables. To keep to the traditional terminology, P is the definition's definiendum and ϕ is its definiens. Of course, Δ as written is only the form of a definition, for we have not said what conduct with the definiendum P it authorizes.[3] To a first approximation, though, the point of advancing Δ is to stipulate that whatever the background facts may be, an object should count as a P just in case it satisfies ϕ. By this act we hope to explain how the Ps are delineated in any given fact-situation.[4] And to explain this is, nearly enough, to explain what the new predicate means.

III. Three Kinds of Definition

Here are some sample definitions to get us started. Bear in mind that everything other than the definiendum P has its meaning given in advance:

(1) $\underline{x}$ is P $=_{df}$ $\underline{x}$ is A or $\underline{x}$ bears E to some A
(2) $\underline{x}$ is P $=_{df}$ $\underline{x}$ is A or $\underline{x}$ bears E to some P
(3) $\underline{x}$ is P $=_{df}$ $\underline{x}$ is A or $\underline{x}$ bears E to some non-P.

Two differences exist among these definitions. The first is that in (1), P appears on the left hand side only, whereas (2) and (3) have P on the right hand side as well. The second is that P appears positively in (2)'s definiens, while in (3)'s it appears negatively.[5] To have words for definitions like these, let's call them noncircular, positive circular, and negative circular, respectively.[6] So

(4) x is P $=_{df}$ x bears E to all As

is a noncircular definition;

(5) x is P $=_{df}$ x bears E to some P or to some A

is positive circular; and

(6) x is P $=_{df}$ no Ps bearing E to x are As

is negative circular. Some of the above are indeed bewildering. The question will be what sense, if any, can still be made of them.

IV. Noncircular Definitions and (D), the Determination Scheme

Not at all bewildering are noncircular definitions, or ordinary explicit definitions of the kind found in logic primers. Their intuitive comprehensibility has an objective basis, for these definitions demonstrably fulfill their promise of showing how to find the definiendum's extension in any given situation or world. The procedure is simple: When a world w is given, we learn what exists in w, and the extensions there of the predicates we understand (including every predicate in ϕ). From this information, Tarski's definition of truth shows how to calculate ϕ's extension in w. And since ϕ is definitionally equivalent to P, the objects in that extension are the ones we should take to satisfy P in w:

(D) **x** satisfies P in w iff **x** belongs to ϕ's extension in w.

(Why do we say "**x** belongs to ϕ's extension" rather than just "**x** satisfies ϕ"? Because (D) instructs us to collect all objects satisfying ϕ before considering what objects P might be true of. Contrast interpretation scheme (E) below.)

V. Circular Definitions and (E), the Equivalence Scheme

Such a procedure for figuring P's extension is not available, though, if Δ is circular. As before, when w is given we learn the extensions in w of every predicate we

understand. The difference is that we do not thereby learn the extensions of all the predicates in ϕ; for ϕ contains P and we do not understand P as yet. This logical difficulty is reflected in the traditional "rule of definition" that the "definiendum must not appear in the definiens."[7] "Definitions which violate this rule are...circular," and

> must be rejected as definitions because they do not explain the meaning of the definiendum: a person who did not already understand the definiendum could not understand the definiens.[8]

There is something right in this criticism: if the definiendum is part of the definiens, then the definiens cannot be fully grasped until the definiendum is fully grasped.[9] But it is a further claim that this renders the definiens useless in an explanation of the definiendum's meaning.[10] Before examining this further claim, let's notice some reasons for wondering whether circular definitions can really be as irredeemable as alleged.

Reading traditional criticisms like the above, one would think that definitions like (2) gave no semantical guidance whatever. This is just untrue. From (2) it follows that the Ps include every A, everything bearing E to an A, everything bearing E to anything bearing E to a A, and so on indefinitely. Information about what is not a P can also be extracted from (2), namely that nothing is a P which is neither an A nor in the domain of the E relation. Related to this, the objector's distrust of circular definitions conflicts with the actual practice of logicians, who use definitions like (2) all the time:[11]

(7) $\underline{x}$ is a number $=_{df}$
$\underline{x}$ is zero or $\underline{x}$ is some number's successor;[12]

(8) $\underline{x}$ is a sentence $=_{df}$
$\underline{x}$ is an atom or $\underline{x}$ is the result of negating or conjoining sentences;

(9) $\underline{x}$ is a theorem $=_{df}$
$\underline{x}$ is an axiom or $\underline{x}$ follows from theorems by modus ponens.

Either logicians are very confused, then, or the critic is overlooking something that logicians have seen. Which is it?

Lurking just in the background here is an issue raised in section II, the issue of how a definition $P\underline{x} =_{df} \phi(\underline{x})$ determines its definiendum's meaning. By far the most usual theory sees P as inheriting its meaning from ϕ. Seen from this perspective, circular

definitions are indeed objectionable. Before it can bestow its meaning on P, ϕ must already have a meaning; and how can it if it contains P as a part? That said, there may be other ways for a definition to fix P's meaning than by furnishing a chunk of language to which that meaning already attaches. This is the possibility that the critic is overlooking.

What other ways can there be, though, for a predicate's definition to settle what it means? Perhaps P's meaning flows, not directly from ϕ's meaning, but from the requirement that these two be (in relevant respects) the same. Bracketing worries about the application-conditional approach to meaning, this amounts to the equivalence scheme

(E) **x** satisfies P in w iff **x** satisfies ϕ in w.

Although technically equivalent to (D), (E) involves a crucial change in perspective: we no longer expect to arrive at P's extension on the basis of ϕ's extension, for if Δ is circular then this is not possible. Instead the proposal is that P's extension should be a set **P** such that when P is taken to stand for **P**, ϕ and P emerge with the same extension. Writing $\phi_{\underline{w}}(\mathbf{P})$ for the extension ϕ receives in w on the hypothesis that P stands for **P**, we can put this in the form of an equation: $\mathbf{P} = \phi_{\underline{w}}(\mathbf{P})$. Whatever solves this equation will be also be said to solve $P\underline{x} =_{df} \phi(\underline{x})$ (in the relevant world). So the proposal is that P's extension should be a solution of its definition.[13]

VI. Positive Circular Definitions and (F), the Forcing Scheme

Whether a definition is circular or not, to ask after its solutions makes perfect sense. However the plural here should give us pause. Definitions are supposed to fix their definienda's meanings uniquely, but circular definitions are seldom uniquely solvable. Negative circular definitions, we'll see, need not be solvable at all; and positive circular definitions, the subject of the present section, tend to be multiply solvable. This is well illustrated by (7), the standard inductive definition of "number." According to the equivalence scheme (E), all that (7) tells us about "number"'s extension is that it should be a set consisting exactly of 0 and its members' successors. But then there can be no objection to an eccentric who concludes from (7) that the "numbers" are the integers, or the rationals, or the reals! For each of these sets contains exactly 0 and the successors of its members.[14]

Doesn't this justify the tradition's rejection of positive circular definitions as logically defective? No it does not. Which rule the words $P\underline{x} =_{df} \phi(\underline{x})$ express is a function not just of the words themselves, but of the interpretation scheme employed. Originally our scheme was (D), which read $P\underline{x} =_{df} \phi(\underline{x})$ as instructing us to

find the things that ϕ applies to, and apply P to them.

When that proved unable to cope with circular definitions we tried the equivalence scheme (E), which read into Δ a different rule:

use P and ϕ in application to the <u>same</u> things.

Now that (E) too has run into trouble, rather than blaming the definitions it seems more natural to blame the scheme.

To go by (E), the only constraint a definition imposes on the use of its definiendum is that P and ϕ should agree in their extensions. But definitions surely demand more of their adherents.

Which objects do I resolve to classify as Ps, when I accept a definition $P\underline{x} =_{df} \phi(\underline{x})$? Not any objects I like, subject only to the requirement that the Ps and the ϕs come out the same; rather the objects that that requirement <u>forces</u> me to treat as Ps. But the only things that I am forced to see as Ps, on pain of having different Ps than ϕs, are the members of Δ's smallest solution. Thus (E) gives way to the <u>forcing</u> scheme

(F) **x** satisfies P in <u>w</u> iff **x** belongs to Δ's least solution in <u>w</u>.

Part of the new scheme's attraction is that, as interpreted by (F), each positive circular definition assigns its definiendum a unique extension in every world. This is because the positive circular definitions are exactly the ones more commonly labeled <u>inductive</u>, and each inductive definition is known to possess a unique least solution. Despite our early misgivings, then, positive circular definitions turn out to be fully as good as explicit definitions at fixing their definienda's meanings.

VII. Negative Circular Definitions and (G), the Grounding Scheme

That leaves definitions like (3), where the definiendum P occurs negatively in the definiens. These are in a good sense opposite to inductive definitions, so let's call them antiinductive.

Antiinductive definitions have hardly been discussed by philosophers, maybe because it seemed there was nothing useful to say about them. Unlike explicit definitions, which are always uniquely solvable, and inductive definitions, which always have a unique least solution, negative circular definitions need not be solvable at all. Even where solutions exist, moreover, they may well be incommensurable in the sense that none is included in all the rest. For an example of an unsolvable definition, consider

(10) x is a snarf $=_{df}$
x is Dan Quayle or x is a non-snarf.

Suppose for contradiction that (10) has a solution, the snarfs. Either Jack Kennedy is a snarf or he is not. On the first hypothesis, either Quayle is Jack Kennedy, which he is not, or Kennedy is a non-snarf, which is contrary to assumption. Yet if Kennedy is not a snarf, then he meets the conditions for snarfhood and must be accounted a snarf after all. This shows that there is no possibility of extensional agreement between (10)'s definiendum and its definiens. The problem is slightly different with definition

(11) x is an anteger $=_{df}$
x is 0 or x is an integer similar to some non-anteger

("x is similar to y" means "x and y are equidistant from zero"). Unlike (10), (11) has solutions: the nonnegative integers {0,1,2,3,...}, for example, or the nonpositive integers, or indeed any set obtaining by choosing one of n and –n as n ranges from 0 forward. These solutions are incommensurable, though, so (11) has no least solution of the sort guaranteed above.

Intractable as antiinductive definitions can be, they do sometimes induce naively acceptable extensions. Take for instance

(12) x is a nember $=_{df}$
x is 0 or x is a number succeeding some non-nember.

Apart from 0, only positive integers can be nembers; so let us start with 1 and work up. To be a nember, 1 would have to succeed a non-nember, which since 0 is a nember it does not. So 1 is not a nember. But then 2 succeeds a non-nember, which makes 2 a nember. By parallel reasoning we see that 3 is a non-nember, 4 is a nember, and so on. Therefore the nembers are precisely the even numbers. This is an intuitive argument but it is surely correct. Another and harder question is why it is correct, that is, what general principle operates to make the even numbers the proper extension?

Now, it may seem as though we have an explanation of this. The even numbers constitute (12)'s least solution, hence they are the objects that the forcing scheme would have us regard as nembers. As we'll see, though, it is something of a fluke that forcing steers us right in this case. Outside the domain of explicit and inductive definitions, it can lead to distinctly counterintuitive results.[15] Even worse, it can fail to give guidance at all.

Actually the second of these two problems has been encountered already. The "snarfs," according to (F), are the things in (10)'s least solution, that is, the things in the smallest set **P** that contains Quayle along with everything that **P** does not contain. But then if there are no such sets (and there are not), (F) leaves us hanging: nothing gets classified either way. This sits ill with the fact that intuitively, Quayle is as clear a case of a snarf as there could be. Now for an example where (F) guides us, all right, but in the wrong direction.[16] Definition

(13) x is ploofy $=_{df}$
at least one thing is ploofy or x is not ploofy,

admits of only one solution: the set of everything whatsoever. (Unless something is ploofy a solution is impossible, for the reader will be ploofy iff she is not ploofy. But if something is ploofy, then the definiens holds for all values of x.) According to (F), then, everything whatsoever should be considered ploofy. This is puzzling since naively (13) offers no grounds for regarding anything as ploofy. Take Ross Perot as our example. For Perot to merit classification as ploofy, he must first satisfy (13)'s definiens ϕ. Because ϕ's right disjunct would require him not to be ploofy, Perot's only chance at ploofiness is to satisfy ϕ's left disjunct, that is, for there to be ploofies already. Extrapolating from this perfectly typical case, nothing deserves to be counted ploofy unless something has already been counted ploofy. How then does anything get to be counted ploofy in the first place? No answer is

possible. There is no way of grounding an attribution of ploofiness to Perot or to anything else.[17]

From these and similar examples it seems that (F) cannot be the whole story about extension-determination. But the examples suggest more than that, for there is a definite pattern to (F)'s lapses. On the one hand, it refuses to count Quayle a snarf, not because Quayle fails to satisfy (12)'s definiens, but because (12)'s definiens and its definiendum cannot be brought into extensional agreement. On the other hand, (F) endorses the hypothesis of universal ploofiness, not because every object demonstrably satisfies (13)'s definiens, but because it represents the only way of keeping (13)'s definiens and definiendum extensionally on a par. The moral is that extensional equivalence, however desirable, must not be pursued at the expense of grounding: the requirement that P should be applied to an object when, and only when, that object has shown itself to satisfy ϕ.[18]

More needs to be said about grounding, but we already have a foot in the door. Remember that it is only when the definiens contains P negatively that the forcing scheme yields unintuitive extensions. This tells us to aim for an interpretation scheme (G) that, just as (F) agreed with (E) on explicit definitions while improving on it elsewhere, agrees with (F) on explicit and inductive definitions while outdoing it on other definitions:

(G) **x** satisfies P in $\underline{w}$ iff **x** is Δ-grounded in $\underline{w}$.

This scheme will be developed in stages. After first giving an account of Δ-grounding appropriate to explicit, inductive and antiinductive definitions, we will find that the natural extension of this account to the one remaining case (definitions such that P has both positive and negative occurrences in the definiens) does not quite work. Luckily the account that does work can be extended backwards to the first three cases in an intuitively satisfying way. That done we will have shown how to interpret all definitions $P\underline{x} =_{df} \phi(\underline{x})$, with no restriction whatever on the definiens.

VIII. Simple Grounding

To begin we re-present Tarski's theory of satisfaction as a theory of grounding for explicit definitions. Assume that we've been given a universe **U** and the extension **A** of every predicate A in some first-order language. Then Tarski's rules show how to determine

which sequences **s** -- which functions from the language's variables into the universe -- satisfy which formulas; or more colloquially, how to tell whether a formula θ is true or false of the objects **s** assigns to its variables. Writing T(θ,**s**) for the first possibility, and F(θ,**s**) for the second, the rules are these:[19]

(AT) $\mathbf{s}(\underline{x}) \in \mathbf{A} \Rightarrow T(A\underline{x}, \mathbf{s})$ (AF) $\mathbf{s}(\underline{x}) \notin \mathbf{A} \Rightarrow F(A\underline{x},\mathbf{s})$

($\neg$ T) $F(\psi,\mathbf{s}) \Rightarrow T(\neg\psi,\mathbf{s})$ ($\neg$ F) $T(\psi,\mathbf{s}) \Rightarrow F(\neg\psi,\mathbf{s})$

($\wedge$T) $T(\psi,\mathbf{s})$ and $T(\chi,\mathbf{s}) \Rightarrow T(\psi\wedge\chi,\mathbf{s})$ ($\wedge$F) $F(\psi,\mathbf{s})$ or $F(\chi,\mathbf{s}) \Rightarrow F(\psi\wedge\chi,\mathbf{s})$

($\forall$T) $T(\psi,\mathbf{s}')$ for all $\mathbf{s}' \approx_{\underline{x}} \mathbf{s} \Rightarrow T(\forall\underline{x}\psi,\mathbf{s})$ ($\forall$F) $F(\psi,\mathbf{s}')$ for some $\mathbf{s}' \approx_{\underline{x}} \mathbf{s} \Rightarrow F(\forall\underline{x}\psi,\mathbf{s})$

These rules taken together are called (A)-($\forall$). Formula θ is said to be true (false) of $\mathbf{x}_1,...,\mathbf{x}_n$ iff (A)-($\forall$) prove T(θ,**s**) for some sequence **s** assigning $\mathbf{x}_1,...,\mathbf{x}_n$ to θ's free variables.[20] Context permitting we write $T(\theta,\mathbf{x}_1,....,\mathbf{x}_n)$ instead of T(θ,**s**); this allows us to say that θ is true of **x**, or equivalently that **x** satisfies θ, iff T(θ,**x**) is obtainable by the stated rules.

Explicit Definitions

Take an ordinary explicit definition $P\underline{x} =_{df} \phi(\underline{x})$. Because P has no occurrences in ϕ, ϕ's extension is determined by (A)-($\forall$) in advance of any information about P's extension. As an obvious corollary, for each **x** in ϕ's extension, (A)-($\forall$) show **x** to be a ϕ without anywhere assuming that **x** is a P. This is just the idea of grounding, so we define: **x** is Δ-grounded iff rules (A)-($\forall$) prove T(ϕ,**x**). Accordingly (G), which says that **x** satisfies P iff **x** is Δ-grounded, acquires the following more particular meaning in connection with explicit definitions:[21]

(G_E) **x** satisfies P iff (A) - ($\forall$) prove T(ϕ,**x**).

Because the set of **x** meeting the latter condition is ϕ's extension, all of our schemes (E), (F) and (G) agree on the interpretation of explicit definitions.

Inductive Definitions

How do we explain inductive definitions, say the standard definition of "number," to an inductive novice?[22] First we might tell him to look for things that satisfy "number"'s

definiens -- "x is 0 or x is the successor of some number" -- no matter what objects are conceived as numbers. From this he deduces that 0, at least, is a number. Next he must learn to bring his current harvest of numbers to bear on the identification of new numbers. Since 0 is a number, "x is 0 or x is the successor of a number" is satisfied, at least, by 0 and 1; since 0 and 1 are numbers, it is satisfied, at least, by 0, 1, and 2; and so on until the process exhausts itself in the definiendum's intended extension. As a matter of fact, the novice is now told, this route to the desired extension is available for all inductive definitions, not just the definition of "number." Objects that satisfy P's definiens ϕ regardless of how P is interpreted may be thrown into P's extension straightaway, whereupon further objects are seen to satisfy ϕ, and so on indefinitely. Repeating this procedure as necessary yields the set inductively defined by $P\underline{x} =_{df} \phi(\underline{x})$.

What interests us in this story is that nothing has been called a P unless it at some point earned that title by showing itself to satisfy the definiens. This is exactly the idea of groundedness, so let us explicitly note the rules involved: they are Tarski's original rules (A)-(∀) plus a rule

(ΔT) $\quad T(\phi,\mathbf{x}) \Rightarrow T(P,\mathbf{x})$

telling us to add **x** to P's extension should the definiens prove true of it. This leads us to call **x** Δ-grounded iff $T(\phi,\mathbf{x})$ is provable using (A)-(∀) and (ΔT); whereupon (G) takes on the meaning

(G_I) $\quad$ **x** satisfies P iff (A)-(∀) and (ΔT) prove $T(\phi,\mathbf{x})$.

But, the set of **x** satisfying (G_I)'s right hand side is known to be Δ's least solution. So when (F) says that P is true of the objects in Δ's least solution, and (G) says that it is true of the Δ-grounded objects, they are saying the same thing.

Antiinductive Definitions

At least, they are saying the same thing if Δ is an inductive definition. Applied to other definitions, we saw, (F) loses its head entirely, leaving intuitively grounded items (remember Quayle) out of P's extension, and making P true of intuitively ungrounded items (remember Perot).

However we have not yet examined what groundedness comes to in the context of antiinductive definitions.[23] Rules (A)-(∀) and (ΔT) are not enough, for they offer no way of showing that P is false of anything, and it is characteristic of antiinductive definitions that ϕ may be true of **x** because P is false of some other thing. This is the case for example with

(12) x is a nember $=_{df}$

x=0 or x is a number succeeding some non-nember.

To show that ϕ is true of 2, we need the information that P is false of 1. And using (A)-(∀) and (ΔT) alone, that information is unavailable. -- Well, how does one show that 1 is not a nember? The reasoning used above was

(a) ϕ is true of 0; so
(b) "nember" is true of 0; so
(c) ϕ is false of 1; so
(d) "nember" is false of 1.

The first step of this reasoning, from (a) to (b), is licensed by

(ΔT) $T(\phi,\mathbf{x}) \Rightarrow T(P,\mathbf{x})$

but the step from (c) to (d) requires a complementary rule

(ΔF) $F(\phi,\mathbf{x}) \Rightarrow F(P,\mathbf{x})$.

This is the only new rule needed to show that the nembers are exactly the even numbers. In fact it is the only new rule needed to work with antiinductive definitions generally. So let's define **x** as Δ-grounded iff (A)-(∀), (ΔT) and (ΔF) prove $T(\phi,\mathbf{x})$, which makes

(G_A) **x** satisfies P iff (A)-(∀), (ΔT) and (ΔF) prove $T(\phi,\mathbf{x})$

the appropriate version of (G) in antiinductive contexts.

Three versions of the grounding scheme have been considered, one for each of our three types of definition: explicit, inductive, and antiinductive. Actually though it is possible to interpret all of these definitions in a uniform way. For notice two things. First, if Δ is explicit, then neither (ΔT) and (ΔF) can contribute to a proof of T(ϕ,**x**). Second, if Δ is inductive, then (ΔF) cannot contribute to such a proof. This means that instead of using (G_E) for explicit definitions, (G_I) for inductive definitions, and (G_A) for antiinductive definitions, we can use (G_A) across the board to the same effect. This is not the only simplification possible. Rules (A)-($\forall$), (ΔT) and (ΔF) prove T(ϕ,**x**) iff they prove T(P,**x**); so, rather than letting **x** satisfy P iff these rules prove T(ϕ,**x**), we may as well adopt the simple grounding scheme

(G_S) **x** satisfies P iff (A)-($\forall$), (ΔT) and (ΔF) prove T(P,**x**).

Thinking of (A)-($\forall$), (ΔT) and (ΔF) as the simple rules, the scheme becomes this: P is to be counted true of **x** iff it is simply provable that P is true of **x**.

IX. Reflective Grounding

Whether Δ is explicit, inductive or antiinductive, the Δ-grounded objects are the ones that can be shown to satisfy P using the simple rules. But there is a kind of definition we have not considered:

(14) x is sheec $=_{df}$
someone sheec applauds x and someone unsheec derides x.

Obviously (14) is not explicit, but "sheec"'s second occurrence in the definiens prevents it from being inductive, while the first occurrence prevents it from being antiinductive. Definitions like this will be called coinductive[24] to reflect the fact that P appears positively and negatively in the definiens.

Now I will give two reasons for thinking that the simple rules are not quite right: one has to do with their treatment of inductive definitions, the other with their treatment of coinductive definitions. Both reasons trace back ultimately to the fact that the simple rules offer no way of reflecting on the grounding process and incorporating the results of that reflection back into the process.

Suppose first that Δ is inductive. All sides agree that the simple rules prove T(P,**x**) for every **x** that P is intuitively true of, so the problem can only be that they fail to establish F(P,**x**) for every **x** that P is intuitively false of. This problem arises even with as familiar a definition as (7), the standard definition of "number." Using (A)-(∀), (ΔT) and (ΔF)[25], the things we can show to be numbers are 0, 1, 2, 3, ... What we cannot show is that nothing else is a number. Should **k** be a negative integer, for instance, then F(ϕ,**k**) can be obtained only by inferring it from F(ϕ,**k**–1) using (ΔF), while F(ϕ,**k**–1) must itself be obtained from F(ϕ,**k**–2) and so on indefinitely. So while **k**'s claim to numberhood cannot be proved, it is not refutable either.

So long as Δ is inductive, this failure to identify all non-Ps is not a matter of real concern. For if P is positive in ϕ, the fact that such and such things fall outside of P's extension cannot qualify anything to belong to ϕ's extension. (This is how the simple rules manage to identify the intuitively correct extension despite underestimating the class of non-Ps.) When we advance to the coinductive realm, however, our luck runs out, for here it can happen that certain items satisfy the definiendum only because other items fail to satisfy it. Look for example at definition

(15) x is a noomber $=_{df}$
x is 0 or x and –x succeed a noomber and a non-noomber respectively.

0 is a noomber because it satisfies ϕ outright. Therefore 0's successor 1 is a noomber provided that –1 succeeds a non-noomber. And so it does, for –2's claim to noomberhood is demonstrably ungroundable: it can qualify for the title of noomber only if –3 has qualified beforehand, only if –4 has qualified before that, only if.... Similar arguments reveal each nonnegative integer to be a noomber while showing that nothing else is a noomber. Next consider

(16) x is awd $=_{df}$
x is identical to an awd number or it succeeds a non-awd number.

0 cannot qualify as awd through ϕ's second disjunct, for it does not succeed any number.[26] Nor does it qualify as awd through ϕ's first disjunct, as this would require it to be awd already, contrary to grounding. So, 0 is not awd, which makes 1 the successor of

a non-awd number and therefore awd. Similar arguments show that 2 is not awd, 3 is awd, 4 is not, and in general that the awd numbers are exactly the odd numbers.

So far, so good, except that the simple rules offer no way of reaching these results. This is because they recognize one route only to the conclusion that P is false of **x**, namely inferring it from the fact that ϕ is false of **x**; and because this route becomes viciously circular in the cases under discussion. To exclude 0 from the set of awd numbers, for example, we would first need to know that it was neither an awd number nor the successor of a non-awd number.

How does it transpire that negative integers are not numbers, or that even numbers are not awd? In all such cases the argument that **x** falsifies P[27] is not that it falsifies ϕ, but that **x**'s claim to satisfy ϕ is demonstrably groundless. This is the argument we try to develop.

By a hypothesis let's mean a set of claims to the effect that such and such formulae are true (false) of such and such objects. Object **x** will be called ungroundable iff T(ϕ,**x**) cannot find a place even in the most inclusive hypothesis compatible with current information. Here is how that theory is constructed. Since the only rule suspected of underproducing is (ΔF), let's start by collecting all claims of the type it proves with the slightest chance of being right, that is, all claims F(P,**s**) such that T(P,**s**) is not a part of current information. Next let's subject this collection to all remaining rules (A)-(∀) and (ΔT). The result is the hypothesis we want, the set of all claims with the slightest chance of being right. Formally, a set Θ of premises makes **x** ungroundable iff T(ϕ,**x**) is not provable using (A)-(∀) and (ΔT) from the set of F(P,**s**) such that T(P,**s**) ∉Θ. Given a Θ of this kind, the reflection rule allows us to infer F(P,**x**):

(ΔR) Θ ⇒ F(P,**x**).....where Θ makes **x** ungroundable.

Here is the same thing in ordinary language: when you know enough to refute all possible proofs of **x**'s claim to satisfy the definiens, you may conclude that the definiendum is false of **x**.

Where does this leave us? Recall that (A)-(∀), (ΔT) and (ΔF) were defined as the simple rules, and that **x** was called simply Δ-grounded iff these rules proved P to be true of **x**. Now let (A)-(∀), (ΔT) and (ΔR) be the reflective rules, and let **x** be reflectively Δ-

grounded iff the revised rules prove P to be true of **x**. As you might expect, both kinds of grounding come to the same if Δ is explicit, inductive, or antiinductive (see the appendix, Prop.5 and the remarks following). And since reflective grounding improves on simple grounding in the area of coinductive definitions, we may as well make it our official notion of grounding for all definitions. By this route we arrive at last at an all purpose interpretation scheme. No matter what kind of definition Δ may be, it instructs us to use its definiendum according to the following scheme:[28]

(G_R) **x** satisfies P iff (A)-(∀), (ΔT) and (ΔR) prove T(P,**x**).

Now we apply this scheme to two bits of unfinished business: inconsistent definitions generally, and an inconsistent definition of truth.

X. Inconsistent Definitions

The project was to make sense of inconsistent definitions, or definitions placing irreconcilable obligations on those adopting them. I claim that a definition demands its adherents to use P in accordance with (E), (F) and (G).[29] This yields the analysis

Δ is consistent iff (E), (F) and (G) are jointly satisfiable; otherwise inconsistent.[30]

Equivalently, since the Δ-grounded objects are the only set satisfying (G), Δ is consistent iff the set of Δ-grounded objects is at the same time Δ's least solution.

How does the analysis deal with our four types of definition? Explicit and inductive definitions are always consistent because, first, they always have a least solution, and second, that least solution is always the set of Δ-grounded objects. The surprising thing is that other definitions can be consistent as well. So although definition (12) of "nember" is antiinductive, and definition (16) of "awd" is coinductive, both of them ground exactly the members of their least solutions.

When a definition is not consistent, this will be for one of two reasons: either it lacks a least solution, or the one it has is at variance with the set of Δ-grounded objects. These possibilities are illustrated by (10) and (13):

(10) x is a snarf $=_{df}$

x is Dan Quayle or x is a non-snarf.

The reflective rules show "snarf" to be true of Quayle and that is all. But when we interpret "snarf" in (10)'s definiens as {Quayle}, the definiens becomes tautologous and therefore satisfiable by objects other than Quayle. This shows that the set of (10)-grounded objects is not a solution; and in fact (10) is absolutely unsolvable. Now consider

(13) x is ploofy $=_{df}$

at least one thing is ploofy is ploofy or x is not ploofy

Definition (13) has a unique solution, viz. the entire universe. But since the reflective rules do not show anything to satisfy "ploofy," this solution is as far from the set of (13)-grounded objects as it could be.

Assessing a definition for consistency might seem a complicated affair. First we find the set of Δ-grounded objects, then we check that it is a solution, and lastly we make sure there are no smaller solutions. Happily there is a simpler method that stays within the grounding process itself. For that process turns up not one but two sets of interest:

$\Gamma_\Delta = \{\mathbf{x} \mid$ the reflective rules prove $T(P,\mathbf{x})\}$,

the set of things that clearly belong in P's extension, and

$\Gamma^\Delta = \{\mathbf{x} \mid$ the reflective rules do not prove $F(P,\mathbf{x})\}$,

the set of things that not clearly belonging outside P's extension. And whether Δ is consistent can be judged from the relation between these two sets: it is consistent if $\Gamma_\Delta = \Gamma^\Delta$, and inconsistent if $\Gamma_\Delta \neq \Gamma^\Delta$ (see the appendix, Prop. 8). Thus the grounding process gives us a sort of idealized decision procedure for the consistency of a definition. For instance, if Δ is

(11) x is an anteger $=_{df}$

x is 0 or x is an integer similar to some non-anteger,

then $\Gamma_\Delta = \{0\} \neq \Gamma^\Delta$ = the set of all integers, reflecting the fact that (11) is inconsistent. Definitions (10) and (13) are also inconsistent by this test, but (12) and (16) come out consistent. All of this agrees with the analysis given at the beginning of the section.

Now, someone could question our treatment of a definition like (11) as inconsistent. After all, what the grounding rules tell us about "anteger" is just that it is true of 0 and false of whatever is not an integer. And this might seem to leave the nonzero integers' status harmlessly undecided. Really the situation is rather worse, though, for we continue to be bound by the equivalence and forcing requirements.[31] Together these make the following reasoning inescapable. By grounding, nothing satisfies "anteger" unless the rules show it to do so. They do not show 7 to satisfy "anteger," so 7 is not an anteger. But then –7 is a non-anteger by the same reasoning; and if neither is an anteger, then each becomes similar to a non-anteger (the other) and so must be counted an anteger to preserve equivalence! Even more confusingly, no sooner do we count them antegers than the basis for this move (that each is similar to a non-anteger) evaporates and we must revert to our original position that they are not antegers. In this way we get caught up in a cycle of semantic reversals, with each reclassification of the nonzero integers immediately calling forth its opposite.[32]

So it is not for no reason that definitions like that of "anteger" are called inconsistent. However we behave with P, we are shirking a semantic obligation, and if we try to fulfill that obligation we find ourselves guilty of some new violation. The predicament is somewhat akin to the conflicts of moral obligation discussed in the ethics literature; for instance, promising your parents you will observe Yom Kippur and your friend you will go stunt flying as soon as she gets her plane fixed. But notice an important difference. Where the usual moral examples involve coexisting imperatives such that complying with one means defying the other, in the semantical case I create the imperative I defy by complying with its competitor. Thus the situation is rather as though "park only in empty spaces" were a bona fide moral law. Whatever parking space I choose is thereby non-empty and so, according to the law, to be vacated. Likewise whatever action I take with P -- pronouncing it true of **x** or not -- gives ϕ the opposite relation to **x**, which forces me as a subscriber to $P\underline{x} =_{df} \phi(\underline{x})$ to take my action back.

XI. Truth

Sound familiar? Substitute "true" for P and "this very sentence is untrue" for **x**, and you get the paradox of the Liar: whatever action I take with "true" -- applying it to the Liar sentence or withholding it -- forces me immediately to reverse myself and take the opposite line. Now I will propose a <u>definition</u> of truth that explains how such a situation arises.[33] Because the definition I favor is not consistent, I will compare it with a consistent definition that is otherwise as similar as possible. That way I can explain why my definition strikes me as more "correct," that is, more in accord with the truth-predicate's ordinary meaning.

By far the most important paper on truth since Tarski is Kripke's "Outline of a Theory of Truth." Kripke does not actually define "true" in the sense we have been concerned with here. But the theory he gives[34] is exactly the one you would get if you took the following simultaneous definition of "true" and "false" and interpreted it according to the procedures of this paper:[35]

ϕ is true $=_{df}$ $\phi = \ulcorner Ra \urcorner$ and a's referent belongs to R's extension,
or $\phi = \ulcorner \sim\psi \urcorner$ and ψ is false
or $\phi = \ulcorner \psi \,\&\, \chi \urcorner$ and both ψ and χ are true
or $\phi = \ulcorner \forall \underline{x}\, \psi(\underline{x}) \urcorner$ and all its instances are true
or $\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is true

ϕ is false $=_{df}$ $\phi = \ulcorner Ra \urcorner$ and a's referent belongs to R's antiextension,
or $\phi = \ulcorner \sim\psi \urcorner$ and ψ is true
or $\phi = \ulcorner \psi \,\&\, \chi \urcorner$ and either ψ or χ is false
or $\phi = \ulcorner \forall \underline{x}\, \psi(\underline{x}) \urcorner$ and at least one of its instances is false
or $\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is false.

I have only one quarrel with this, and it concerns the very last line: unless I am mistaken, if ψ is <u>not true</u>, then a sentence saying it <u>is</u> true ought to be considered false. So, I would replace Kripke's "$\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is false" with "$\phi = \ulcorner \psi$ is true$\urcorner$ and ψ is not true." This adjustment, although superficially minute, makes an enormous logical difference: Kripke's definition is inductive and so consistent, whereas mine is coinductive and, although this does not follow from coinductiveness, inconsistent.

How does the proposed adjustment bear on the Liar sentence? With "true" interpreted Kripke's way, the Liar creates a problem only for those who insist on assigning it a truth value; seen as neither true nor false it causes no more harm than any other truth value gap, for instance, "the President of Moosejaw likes maple syrup." But on my interpretation, whoever reckons the Liar neither true nor false obliges herself to count it true after all. More generally, if Kripke's definition is correct then we can meet our obligations regarding the Liar, whereas if mine is correct we cannot: the Liar emerges as the most striking manifestation of "true"'s inconsistency, or better, the inconsistency of the semantic rules that together constitute its meaning.[36]

REFERENCES

A. Gupta: 1982, "Truth and Paradox," Journal of Philosophical Logic **11**, 1-60

A. Gupta & N. Belnap: 1993, The Revision Theory of Truth, MIT Press, Bradford Books, Cambridge MA

S. Kripke: 1975, "Outline of a Theory of Truth," Journal of Philosophy **72**, pp. 690-716

H. Leonard: 1967, Principles of Reasoning: Introduction to Logic, Methodology, and the Theory of Signs, Dover, London

S. Yablo: 1985, "Truth and Reflection," Journal of Philosophical Logic **14**, 297-349

S. Yablo: 1993, "Rules for Truth," in J. Tomberlin, ed., Philosophical Perspectives **7–8**

APPENDIX

For the formal theory, we use an ordinary first-order language L with connectives $\neg$ and $\wedge$ (negation and conjunction) and quantifier $\forall$ (universal generalization). Other connectives and quantifiers should be seen as defined from these in the usual way.[37] L(P) is the language obtained by supplementing L with a new predicate P. A definition of P is something of the form

$$\Delta \qquad P\underline{x} =_{df} \phi(\underline{x}),$$

where $\phi(\underline{x})$ is a formula of L(P). Here is our problem: given a classical model M of L, how do we tell which members of M's universe P is true of and which it is false of? The problem is broken down into four cases, corresponding to the four types of definition:

- explicit (P does not occur in ϕ);
- inductive (P occurs positively in ϕ);
- antiinductive (P occurs negatively in ϕ);
- coinductive (none of the above),

What we'll see is that the equivalence scheme (E) covers the first case; the forcing scheme (F) covers the first two cases; the simple grounding scheme (G_S) covers the first three cases; and the reflective grounding scheme (G_R) covers all definitions.

For each model M of L, and each subset **P** of M's domain, let M**P** be the obvious expansion of M to L(P): the model that interprets P as **P** and everything else as M interprets it. For each formula ψ of L(P) in one free variable, let M**P**[ψ] be the set of all **x** such that M**P** $\models \psi[\mathbf{x}]$, that is, the set of objects satisfying ψ in the expanded model. Then the equivalence scheme (E), according to which the same objects should satisfy P as ϕ, comes down to this: **P** = M**P**[ϕ]. Sets meeting this condition are called solutions of $P\underline{x} =_{df} \phi(\underline{x})$ in model M, or context permitting, just solutions.

> Prop.1: Every explicit definition has a unique solution in every model.
> Proof: Because P does not occur in ϕ, M**P**[ϕ] = M[ϕ] for all choices of **P** . So **P** = M[ϕ] is Δ's unique solution. ¶

With circular definitions, we saw, a unique solution is not guaranteed. However something like Prop.1 holds of positively circular, or inductive, definitions.

Roughly and intuitively, whether P is negative or positive in ϕ turns on whether it is negated at the level of deepest logical form. Thus P is negative in $P\underline{x}\rightarrow Q\underline{x}$ because the latter reduces to $\neg P\underline{x}\vee Q\underline{x}$, yet positive in $\neg (P\underline{x}\rightarrow Q\underline{x})$ because of the equivalence with $P\underline{x}\wedge\neg Q\underline{x}$. Although it is common to draw the distinction syntactically we will take a semantical approach, calling P positive in ϕ iff the larger P's extension is, the larger ϕ's extension is, and negative in ϕ just in case the opposite relation holds:

Def. P is positive in ϕ iff for all M and all $\mathbf{P}\subseteq\mathbf{Q}\subseteq$dom(M), $M\mathbf{P}[\phi]\subseteq M\mathbf{Q}[\phi]$.
P is negative in ϕ iff for all M and all $\mathbf{P}\subseteq\mathbf{Q}\subseteq$dom(M), $M\mathbf{Q}[\phi]\subseteq M\mathbf{P}[\phi]$.

Note that P is trivially positive in any ϕ that does not contain it (and negative too for that matter). Thus it does not suffice to make a definition inductive for P to be positive in ϕ; P must occur positively in ϕ, meaning that P both occurs in ϕ and is positive in it. Likewise Δ is antiinductive iff P occurs in ϕ and is negative in ϕ.

Assume a fixed model M of L. By the jump operator associated with M, we mean the function J taking **S** to $M\mathbf{S}[\phi]$ = the set of things that satisfy ϕ when P is interpreted as **S**. From the definitions it's clear that if Δ is inductive, this operator is monotonic in the sense of preserving inclusion relations. (For all **X** and **Y**, $\mathbf{X}\subseteq\mathbf{Y}\Rightarrow J(\mathbf{X})\subseteq J(\mathbf{Y})$).

Prop.2 Every inductive definition has a least solution in every model.
Proof: Consider the sequence $\mathbf{P}_0=\varnothing$, $\mathbf{P}_\alpha = J(\mathbf{P}_{\alpha-1})$.[38] Since J is monotonic, $\langle\mathbf{P}_\alpha\rangle$ is increasing[39]. For cardinality reasons $\langle\mathbf{P}_\alpha\rangle$ eventually reaches a $\mathbf{P}_\gamma = \mathbf{P}_{\gamma+1}$. Let this be **P**. $\mathbf{P} = J(\mathbf{P}) = M\mathbf{P}[\phi]$, so **P** is a solution. To see that **P** is least, let **Z** be any other solution. By J's monotonicity, $\mathbf{P}_\alpha\subseteq\mathbf{Z}\Rightarrow\mathbf{P}_{\alpha+1}\subseteq J(\mathbf{Z})=\mathbf{Z}$. Transfinite induction shows that $\mathbf{P}_\gamma\subseteq\mathbf{Z}$. ¶

Antiinductive definitions need not be solvable at all, so Prop.2 cannot be extended to them. Notice where the proof breaks down: if Δ is antiinductive, the jump operator J is antimonotonic. (Meaning that it reverses inclusion relations: $\mathbf{X}\subseteq\mathbf{Y}\Rightarrow J(\mathbf{Y})\subseteq J(\mathbf{X})$.)

Def. A semisolution of Δ is a pair of sets **P** and **Q**, **P** a subset of **Q**, such that (i) $\mathbf{P} = M\mathbf{Q}[\phi]$ and (ii) $\mathbf{Q} = M\mathbf{P}[\phi]$.

P and **Q** are Δ's least semisolution iff all semisolutions **X** and **Y** lie between them, that is, **P**⊆**X**⊆**Y**⊆**Q**.

> Prop.3 Every antiinductive definition has a least semisolution.
> Proof: Since J is antimonotonic, H = J∘J is monotonic. Define the sequence $\langle \mathbf{P}_\alpha \rangle$ by $\mathbf{P}_0 = \varnothing$, $\mathbf{P}_\alpha = H(\mathbf{P}_{\alpha-1})$,[40] and the sequence $\langle \mathbf{P}^\alpha \rangle$ by $\mathbf{P}^\alpha = J(\mathbf{P}_\alpha)$. Then $\langle \mathbf{P}_\alpha \rangle$ is increasing and $\langle \mathbf{P}^\alpha \rangle$ is decreasing, with $\mathbf{P}_\alpha$ always a subset of $\mathbf{P}^\alpha$. For cardinality reasons, there exists a γ such that $\mathbf{P}_\gamma = \mathbf{P}_{\gamma+1}$ and $\mathbf{P}^\gamma = \mathbf{P}^{\gamma+1}$. Let $\mathbf{P} = \mathbf{P}_\gamma$ and $\mathbf{Q} = \mathbf{P}^\gamma$. Then $\mathbf{P} = J(J(\mathbf{P})) = J(\mathbf{Q}) = M\mathbf{Q}[\phi]$, and $\mathbf{Q} = J(J(\mathbf{Q})) = J(\mathbf{P}) = M\mathbf{P}[\phi]$. So **P** and **Q** semisolve Δ. Leastness is proved by ordinal induction. ¶

Least semisolutions have been encountered already in another guise: if Δ is anything but coinductive, **P** and **Q** are Δ's least semisolution iff **P** is the set of **x** such that T(P,**x**) is simply provable and **Q** is the set of **x** such that F(P,**x**) is not simply provable. (I will take this for granted in what follows.)

That leaves the fourth case, where P occurs in ϕ but with no particular valence. This can come about only if P makes multiple appearances in ϕ, some in a positive position and others in a negative one.[41] But what is it for an occurrence of P to be positive (negative)? Basically the idea is that increasing that one occurrence's extension, leaving all else the same, will increase (decrease) ϕ's extension. Given an occurrence P_k of P in ϕ, let ϕ_k be result of replacing all other occurrences of P in ϕ with occurrences of some L-predicate C not occurring in ϕ. Then

> Def. P_k is a positive occurrence of P in ϕ iff P is positive in ϕ_k, and a negative occurrence of P in ϕ iff P is negative in ϕ_k.

So P's first occurrence in $P\underline{x} \rightarrow (A\underline{x} \rightarrow P\underline{x})$ is negative, since P is negative in $\phi_1 = P\underline{x} \rightarrow (A\underline{x} \rightarrow C\underline{x})$; but its second occurrence is positive because P is positive in $\phi_2 = C\underline{x} \rightarrow (A\underline{x} \rightarrow P\underline{x})$. P_k is said to have a polarity in its containing formula iff it is either positive in that formula or negative in it.

Lemma: Each occurrence of P in ϕ has a polarity.
Proof: By induction on complexity, the result holds for all formulas ϕ containing P exactly once. Now let ϕ be an arbitrary P-containing formula: P_k has a polarity in ϕ iff P has a polarity in ϕ_k, which it must since ϕ_k contains P exactly once. ¶

That every predicate-occurrence has a polarity allows us to combine our methods for inductive and antiinductive definitions.

Def. Where ϕ is a formula of L(P), ϕ# (ϕ's polarization) is the result of replacing all positive occurrences of P in ϕ with $P_{\#}$, and all negative occurrences of P in ϕ with $P^{\#}$. For M a model of L, $M^{\mathbf{Y}}_{\mathbf{X}}$ is the model of $L(P_{\#},P^{\#})$ that is just like M except in assigning **X** to $P_{\#}$ and **Y** to $P^{\#}$.

Often we write $M^{\mathbf{Y}}_{\mathbf{X}}[\phi]$ for $M^{\mathbf{Y}}_{\mathbf{X}}[\phi\#]$. This means that $M^{\mathbf{Y}}_{\mathbf{X}}[\phi]$ is the set of objects that satisfy ϕ when **X** is assigned to P's positive occurrences in ϕ and **Y** is assigned to its negative ones.

Def. A semisemisolution of Δ is a pair of sets **P** and **Q**, **P** a subset of **Q**, such that (i) $\mathbf{P} = M^{\mathbf{Q}}_{\mathbf{P}}[\phi]$ and (ii) $\mathbf{Q} = M^{\mathbf{P}}_{\mathbf{Q}}[\phi]$.

Prop.5 will show that if Δ is antiinductive, its least semisemisolution **P**, **Q** is its least semisolution as well; if Δ is inductive, **P** is Δ's least solution; and if Δ is explicit, **P** is Δ's unique solution. Therefore all of our propositions so far can be obtained as corollaries of

Prop.4 Every definition has a least semisemisolution.
Proof: First we introduce L, the leap operator. Because $M^{\mathbf{Y}}_{\mathbf{X}}[\phi]$ is monotonic in **X**, for any **Z** the sequence $\mathbf{S}_0 = M^{\mathbf{Z}}_{\varnothing}[\phi]$, $\mathbf{S}_\alpha = M^{\mathbf{Z}}_{\mathbf{S}_{\alpha-1}}[\phi]$ is increasing. For cardinality reasons it reaches a limit $\mathbf{S}_\gamma = \mathbf{S}_{\gamma+1} = M^{\mathbf{Z}}_{\mathbf{S}_\gamma}[\phi]$. This $\mathbf{S}_\gamma$ will be L(**Z**), with the result that (*) $L(\mathbf{Z}) = M^{\mathbf{Z}}_{L(\mathbf{Z})}[\phi]$. Now, since $M^{\mathbf{Y}}_{\mathbf{X}}[\phi]$ varies inversely with **Y**, L is an antimonotonic operator, whence $K = L^{\circ}L$ is monotonic. Define the sequence $\langle\mathbf{P}_\alpha\rangle$ by $\mathbf{P}_0 = \varnothing$, $\mathbf{P}_\beta = K(\mathbf{P}_{\beta-1})$; and define $\langle\mathbf{P}^\alpha\rangle$ by $\mathbf{P}^\beta = L(\mathbf{P}_\beta)$. These sequences are increasing and decreasing respectively, so ultimately they arrive at fixed points **P** and **Q** of the K operator. Since $\mathbf{P} = L(\mathbf{Q})$ and $\mathbf{Q} = L(\mathbf{P})$, it follows from (*) that **P** is $M^{\mathbf{Q}}_{\mathbf{P}}[\phi]$ and **Q** is $M^{\mathbf{P}}_{\mathbf{Q}}[\phi]$. Evidently $\mathbf{P} \subseteq \mathbf{Q}$, so **P** and **Q** semisemisolve the definition. That **P** and **Q** are least follows by induction. ¶

No matter what kind of definition Δ is, Δ's least semisemisolution is the pair consisting of $\Gamma_\Delta = \{\mathbf{x} \mid T(P,\mathbf{x})$ is reflectively provable$\}$ and $\Gamma^\Delta = \{\mathbf{x} \mid F(P,\mathbf{x})$ is not reflectively provable$\}$. So the next proposition says in effect that the reflective rules yield, first, each explicit definition's unique solution; second, each inductive definition's least solution; third, each antiinductive definition's least semisolution; and fourth, each coinductive definition's least semisemisolution.

Prop.5 Let **P**, **Q** be Δ's least semisemisolution. Then

(i) Δ is explicit $\Rightarrow$ **P** = **Q** = Δ's unique solution

(ii) Δ is inductive $\Rightarrow$ **P** = **Q** = Δ's least solution, and

(iii) Δ is antiinductive $\Rightarrow$ **P**, **Q** = Δ's least semisolution

Proof: [Δ explicit] $\mathbf{P} = M^{\mathbf{Q}}_{\mathbf{P}}[\phi] \Rightarrow \mathbf{P} = M[\phi]$ since P does not occur in ϕ. Likewise $\mathbf{Q} = M^{\mathbf{P}}_{\mathbf{Q}}[\phi] \Rightarrow \mathbf{Q} = M[\phi]$. By Prop. 1, $M[\phi]$ is Δ's unique solution. [Δ inductive] Since P has no negative occurrences in ϕ, $\mathbf{P} = M^{\mathbf{Q}}_{\mathbf{P}}[\phi] \Rightarrow \mathbf{P} = M\mathbf{P}[\phi]$ and $\mathbf{Q} = M^{\mathbf{P}}_{\mathbf{Q}}[\phi] \Rightarrow \mathbf{Q} = M\mathbf{Q}[\phi]$. So **P** and **Q** both solve Δ. To see that **P** is Δ's least solution: **X** solves Δ $\Rightarrow$ **X** and **X** semisemisolve $\Delta \Rightarrow \mathbf{P} \subseteq \mathbf{X} \subseteq \mathbf{Q}$ since **P**,**Q** is Δ's least semisemisolution. To see that **P** = **Q**, it's enough to show that $\mathbf{Q} \subseteq \mathbf{P}$. $\mathbf{x} \in \mathbf{Q} \Rightarrow F(P,\mathbf{x})$ is not reflectively provable $\Rightarrow$ **P** does not make **x** ungroundable $\Rightarrow T(\phi,\mathbf{x})$ follows from $\{F(P,\mathbf{z}) \mid \mathbf{z} \notin \mathbf{P}\}$ using (A)-($\forall$) and (ΔT) $\Rightarrow T(\phi,\mathbf{x})$ follows from the null set using (A)-($\forall$) and (ΔT) (since P is positive in ϕ) $\Rightarrow T(P,\mathbf{x})$ follows from the null set using (A)-($\forall$) and (ΔT) $\Rightarrow \mathbf{x} \in \mathbf{P}$. [$\Delta$ antiinductive] Since P has only negative occurrences in ϕ, $\mathbf{P} = M^{\mathbf{Q}}_{\mathbf{P}}[\phi]$ $\Rightarrow \mathbf{P} = M\mathbf{Q}[\phi]$ and $\mathbf{Q} = M^{\mathbf{P}}_{\mathbf{Q}}[\phi] \Rightarrow \mathbf{Q} = M\mathbf{P}[\phi]$. So **P**, **Q** is a semisolution of Δ. For leastness, any other semisolution **X**, **Y** is also a semisemisolution, and **P**, **Q** is by hypothesis least among Δ's semisemisolutions. ¶

Since least semisemisolutions are constructed using the reflective rules, and least solutions and semisolutions are constructed using the simple rules, it follows from Prop. 5 that if Δ is anything but coinductive, the simply Δ-grounded objects are exactly the reflectively Δ-grounded objects. The next proposition adds that whatever type of definition Δ may be, the reflectively Δ-grounded objects include the simply Δ-grounded objects.

Prop.6: For all definitions Δ, simple Δ-groundedness entails reflective Δ-groundedness.

Proof: This is clear from Prop.5 for explicit, inductive and antiinductive definitions, so let Δ be coinductive. The only simple rule that is not also a reflective rules is (ΔF), so

it suffices to show that the set Π_Δ of reflectively provable claims is closed under (ΔF), that is: $F(\phi,\mathbf{x})\in\Pi_\Delta \Rightarrow F(P,\mathbf{x})\in\Pi_\Delta$. Induction on complexity shows that for any ψ and $\mathbf{y}$, $F(\psi,\mathbf{y})\in\Pi_\Delta \Rightarrow T(\psi,\mathbf{y})$ is <u>not</u> a member of the set Π^Δ of claims provable from $\{F(P,\mathbf{z}) \mid T(\phi,\mathbf{z})\notin\Pi_\Delta\}$ using (A)-($\forall$) and (ΔT). Thus $F(\phi,\mathbf{x})\in\Pi_\Delta \Rightarrow T(\phi,\mathbf{x})\notin\Pi^\Delta \Rightarrow$ $T(\phi,\mathbf{x})$ is not provable from $\{F(P,\mathbf{z}) \mid T(\phi,\mathbf{z})\notin\Pi_\Delta\}$ by (A)-($\forall$) and (ΔT) $\Rightarrow \Gamma_\Delta$ makes $\mathbf{x}$ ungroundable $\Rightarrow F(P,\mathbf{x})$ is obtainable from Γ_Δ by (ΔR) $\Rightarrow F(P,\mathbf{x})\in\Gamma_\Delta$. ¶

To finish we verify two assertions about consistency made in the text. The first says that Δ is consistent just in case P has an extension **P** satisfying (E) and (G) -- or what is equivalent, Γ_Δ solves Δ. (Remember that $\Gamma_\Delta = \{\mathbf{x} \mid \mathbf{x}$ is Δ-grounded$\}$ = the set of $\mathbf{x}$ such that $T(P,\mathbf{x})$ is reflectively provable.)

<u>Prop. 7</u> Δ is consistent $\Leftrightarrow$ Γ_Δ solves Δ.
<u>Proof</u>: [$\Rightarrow$] Trivial. [$\Leftarrow$] **P** satisfies (E) $\Rightarrow$ **P** solves Δ. **P** satisfies (G) $\Rightarrow$ **P** = $\Gamma_\Delta \Rightarrow$ **P**, **Q** is Δ's least semisemisolution (where **Q** = Γ^Δ). To see that **P** is Δ's least solution, let **X** be any other solution. Then trivially **X**, **X** semisemisolves Δ. But **P**, **Q** is least among Δ's semisemisolutions, so $\mathbf{P}\subseteq\mathbf{X}\subseteq\mathbf{Q}$. Thus **P** is Δ's least solution $\Rightarrow$ **P** satisfies (E), (F) and (G) $\Rightarrow \Delta$ is consistent. ¶

Second, Δ is consistent iff for every $\mathbf{x}$, exactly one of $T(P,\mathbf{x})$ and $F(P,\mathbf{x})$ is reflectively provable.

<u>Prop.8</u> Δ is consistent $\Leftrightarrow$ $\Gamma_\Delta = \Gamma^\Delta$.
<u>Proof</u>: [$\Leftarrow$] Let **P** = $\Gamma_\Delta = \Gamma^\Delta$. Then since Γ_Δ, Γ^Δ is a semisemisolution, $\mathbf{P} = M^{\mathbf{P}}_{\mathbf{P}}[\phi] = M\mathbf{P}[\phi]$. So **P** = Γ_Δ is a solution, whence Δ is consistent by the last proposition. [$\Rightarrow$] That $\Gamma_\Delta\subseteq\Gamma^\Delta$ is easy, so we show that $\Gamma^\Delta\subseteq\Gamma_\Delta$. Suppose not. Γ^Δ is the set of all $\mathbf{y}$ such that $T(\phi,\mathbf{y})$ can be proved from $\{F(P,\mathbf{x}) \mid \mathbf{x}\notin\Gamma_\Delta\}$ using (A)-($\forall$) and (ΔT). By assumption, some such proofs have $\mathbf{y}\notin\Gamma_\Delta$; among these choose π to be one of shortest length. Since Γ_Δ solves Δ, Γ_Δ is the set of all $\mathbf{y}$ such that $T(\phi,\mathbf{y})$ is provable from $\{T(P,\mathbf{x}) \mid \mathbf{x}\in\Gamma_\Delta\}\cup\{F(P,\mathbf{x}) \mid \mathbf{x}\notin\Gamma_\Delta\}$ using (A)-($\forall$). Thus π can prove $T(\phi,\mathbf{y})$ with $\mathbf{y}\notin\Gamma_\Delta$ only by using (ΔT) at some point to obtain a $T(P,\mathbf{z})$ such that $\mathbf{z}\notin\Gamma_\Delta$. Given the structure of (ΔT) this requires that some proper subproof of π proves $T(\phi,\mathbf{z})$, contrary to our assumption that π was shortest among proofs of this kind. ¶

So, a definition is consistent iff it divides the universe into two parts: the part that the definiendum is true of and the part that it is false of.

---

* This paper was written in a rush, so please forgive the occasional goof-up. Thanks to James Joyce, Ruth Millikan, Leon Porter, Gideon Rosen, and especially Sally Haslanger for help and advice.

[1] Two remarks. Definitions can be either of new words or of words already in use. But even in the latter case it is as though the defined word was new, for the definition must not assume a meaning for it in assigning it a meaning. Only after the definition has done its work do we compare the meaning assigned with the one the word had already, pronouncing the definition correct iff they agree. For this reason I will sometimes call the defined word "new" or "not yet understood" even where an antecedent meaning exists. Second, definitions will seem truth-valuable if one neglects the distinction between defining a word and asserting that the given definition is true to the word's existing meaning. "'Tables' are hereby defined as bird-dogs" is eccentric but not false; "'tables' are correctly defined as bird-dogs" is another matter.

[2] Because our focus will be on predicates, "rules of usage" should be understood as "rules of application."

[3] This job falls to what I will call an interpretation scheme: a way of telling what rule for the use of P is encoded in a string of the form $P\underline{x} =_{df} \phi(\underline{x})$. Several such schemes will be considered below.

[4] As you can see from the last two sentences, I am going to be extremely sloppy about use and mention.

[5] For now I leave the notions of positive and negative at an intuitive level. See the appendix for an exact definition.

[6] This classification is not exhaustive: a fourth category, combining the features of positive and negative circular definitions, will become important later on.

[7] Leonard 1967, p.363.

[8] Leonard 1967, p.364.

[9] Even this may be conceding too much, but let it pass.

[10] Perhaps the situation is as follows: we can partly grasp the definiens without understanding the definiendum at all; this partial grasp can be parlayed into a partial grasp of the definiendum and thereby an improved grasp of the definiens; and so on until we arrive at a full understanding of both.

[11] Some will protest that to the logician, (7) - (9) are no more than shorthands for higher order explicit definitions (on the model of '$\underline{x}$ is a number $=_{df}$ $\underline{x}$ belongs to the smallest set containing 0 and closed under successor'). Arguably though the shoe is on the other foot: ascending to a higher order is just the philosopher's way of calming her conscience about circularity. Left to themselves, logicians take (7) - (9) at face value.

[12] For purposes of this paper, "numbers" are natural numbers.

[13] Notice that (E) agrees with (D) where explicit definitions are concerned. Unless ϕ contains P, the one and only set satisfying (D) in a world is ϕ's extension in that world.

[14] All and only integers, for instance, are either 0 or the successors of integers. The objection that negative integers fail to satisfy the definiens leaves our eccentric unmoved: he replies that –1 succeeds the "number" –2, which succeeds the "number" –3, and so on.

[15] Just to be clear about the shape of the eventual point, the forcing scheme is not wrong to say that Δ instructs us to apply P to the members of Δ's least solution. But Δ also issues other instructions for the use of P; and in cases of conflict, these other instructions take precedence over those urged by the forcing scheme.

[16] For related examples see Gupta 1982 and Gupta & Belnap 1993.

[17] Someone might reply that although we cannot ground an attribution of ploofiness to **x**, still we have an argument for counting it ploofy, viz. that otherwise we are caught up in a paradox. But that kind of argument can be constructed for any conclusion, just by taking a paradox and installing that conclusion as the only escape route.

[18] Earlier we rejected the traditional idea that ϕ's extension must be ascertainable in advance of P's extension. But that idea contained a germ of truth: ϕ's applicability to a particular **x** must be ascertainable without assuming that P applies to **x**.

[19] Sequences are assignments of universe elements to each of L's variables. $\mathbf{s}' \approx_{\underline{x}} \mathbf{s}$ means that $\mathbf{s}'$ is like $\mathbf{s}$, except that $\mathbf{s}'(\underline{x})$ can be any element of the universe. To reduce clutter I omit explicit relativization to a world $\underline{w}$. But strictly speaking (A) should be $(\mathsf{A}_{\underline{w}})$, obtained by replacing **A** with a world-relative extension $\mathbf{A}_{\underline{w}}$, and $(\forall)$ should be $(\forall_{\underline{w}})$, obtained by letting $\mathbf{s}'$ range over sequences assigning $\underline{x}$ a member of $\mathbf{U}_{\underline{w}}$ = $\underline{w}$'s universe.

[20] "Prove" could be misleading since $(\forall)$ is an infinitary rule; I use it anyway.

[21] Again I omit explicit relativization to a world $\underline{w}$. Strictly speaking (G_E) should be: **x** satisfies P in $\underline{w}$ iff $(\mathsf{A}_{\underline{w}})$-$(\forall_{\underline{w}})$ prove $T(\phi,\mathbf{x})$. Similar remarks apply to versions of (G) presented below.

[22] Compare a famous passage in Kripke 1975.

[23] Meaning, definitions $P\underline{x} =_{df} \phi(\underline{x})$ such that P is negative in ϕ.

[24] Mathematicians use "coinductive" not for a special type of definition but a special type of set: a set whose complement can be defined inductively.

[25] Actually (ΔF) is not needed; see Section VIII.

[26] As always, the numbers are the natural numbers.

[27] "**x** falsifies ϕ" is another way of saying that ϕ is false of **x**.

[28] I don't mean to suggest that (G_R) exhausts the obligations a definition imposes on its devotees. On the contrary, (E) and (F) continue to apply. However in cases of conflict (see Section X), (G_R) takes precedence.

[29] Unless otherwise indicated, (G) is (G_R), the reflective grounding scheme.

[30] Two remarks. First, it would be equivalent to call Δ consistent iff some **P** satisfied (E) and (G) alone; any **P** that does that much it is bound to satisfy (F) as well (Prop.7). Second, consistency should really be relative to a world; like truth it is "risky" (Kripke 1975).

[31] Because the forcing requirement follows from the other two (Prop.7), I will stick just to equivalence and grounding.

[32] That these reversals come so naturally might tempt someone to turn the objection on its head: if the existing rules do not allow for flip-flopping, then we need some that do. (See Yablo 1993.)

[33] See also Yablo 1985 and 1993.

[34] Actually I am talking about just one component of Kripke's theory, his construction of the minimal fixed point based on the strong Kleene valuation scheme. But this is the component that has attracted the most attention.

[35] Although we have not discussed simultaneous definitions in so many words, they are handled in the obvious way: each of (ΔT), (ΔF) and (ΔR) becomes two rules, one per definiendum.

[36] Of course, to call these rules inconsistent is only to say that we are not always able to do what they ask. Quite often we can do what they ask, and in these cases definitive semantical classification is possible. Just as the inconsistency of definition (11) doesn't prevent 0 from being a clear case of an anteger, the inconsistency of our definition of "true" doesn't prevent "snow is white" from being a clear case of a truth. But where 0 is the only clear case of an anteger, our definition of truth recognizes infinitely many truths and falsehoods. (Actually it assigns more sentences truth values than Kripke's consistent definition does.)

[37] This proviso is meant to be taken seriously: some of what we do below becomes incorrect if the biconditional, for instance, is taken as basic (see note 41).

[38] $\mathbf{P}_{\alpha-1}$ is to be understood as $\mathbf{P}_\beta$ if $\alpha = \beta+1$, or $\cup_{\beta<\alpha}\mathbf{P}_\beta$ if α is a limit ordinal.

[39] Here and throughout "increasing" is used in the weak sense: $\alpha \leq \beta \Rightarrow \mathbf{P}_\alpha \subseteq \mathbf{P}_\beta$.

[40] As before, $\mathbf{P}_{\alpha\text{-}1}$ is to be understood as $\mathbf{P}_\beta$ if $\alpha = \beta+1$, or $\cup_{\beta<\alpha}\mathbf{P}_\beta$ if α is a limit ordinal.

[41] Suppose that the biconditional had been treated as basic. Then it could have happened that P had no particular valence in ϕ despite occurring only once in it; for instance, P is neither positive nor negative in $P\underline{x} \leftrightarrow Q\underline{x}$. As it is, though, $P\underline{x} \leftrightarrow Q\underline{x}$ abbreviates the conjunction $(P\underline{x} \rightarrow Q\underline{x}) \,\&\, (Q\underline{x} \rightarrow P\underline{x})$. This contains P twice, the first occurrence being negative and the second positive.