Philosophical Perspectives, 7, Language and Logic, 1993

HOP, SKIP AND JUMP: THE AGONISTIC CONCEPTION OF TRUTH[1]

Stephen Yablo
University of Michigan, Ann Arbor

> Is there harm in the contradiction that arises when someone says: "I am lying.—So I am not lying.—So I am lying.—etc."? I mean: does it make our language less usable if in this case, according to the ordinary rules, a proposition yields its contradictory, and vice versa?
> Wittgenstein, *Remarks on the Foundations of Mathematics*

I.

Now and again Wittgenstein hints that our concept of truth is (to use a word) *inconsistent*. What could this mean? Not that the concept is *unsatisfiable*, surely, for in that case no sentences would be true, and some are. Does our truth-concept force inconsistent *conclusions* on us, conclusions to the effect that something both is and is not the case?[2] Wittgenstein's own example tells a different and more convincing story: we *flip-flop* between "I am lying" and "I am not lying," but *without* embracing their contradictory conjunction.[3]

Maybe it is this very flip-flopping, then, the fact that we are "continuously being driven from one decision to the contrary one,"[4] that marks our truth-concept as inconsistent. Such a view looks like mistaking a symptom of the inconsistency for the thing itself. Shouldn't our truth-concept have some *prior* feature to which the flip-flopping could then be attributed? This prior feature, if it existed, would *constitute* truth's inconsistency rather than just manifesting it.[5] But it is not easy to think what the feature could be.

Another Wittgensteinian idea now inserts itself, that a word's meaning, or the concept it expresses, is a matter of the rules that regulate its employment.[6] Such rules are usually conceived as putting conditions on *objects*—as specifying what a thing must be like for the word to apply to it—but it is equally true that they make demands on *subjects*, spelling out what sort of conduct with the word is or is not permitted. This has an interesting and highly relevant consequence, namely that there are *two* ways for a concept to be unsatisfiable. The more usual

way is for it to impose unsatisfiable conditions on *objects*; this is illustrated by the concept of a round square, or a totalitarian democracy. The other, and neglected, alternative is for the concept to make impossible demands on *subjects*: situations arise such that *however* one behaves with the word, one is defying some semantic duty. This is illustrated, it seems to me, by the concept of truth.

II.

How much of our behavior with a predicate P does its associated concept attempt to regulate? Not *all* of it, certainly: whether one uses P on holidays, or under water, is a matter of individual choice. What one *applies* P to, though, is something the concept has views about.

To keep things simple, let's take the regulated behavior to be just the following: saying which objects P is true of, and which it is false of. Assuming that there is a canonical name $\ulcorner\alpha\urcorner$ for each object α, this amounts to assessing atomic sentences $P\ulcorner\alpha\urcorner$ for truth and falsity. So the concept is given by a rule like this:

$$P\ulcorner\alpha\urcorner \text{ is to be counted } \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \text{ iff } \alpha \text{ is } \left\{\begin{matrix}\text{thus and so}\\ \text{thus and such}\end{matrix}\right\}$$

Evaluation of more complex sentences proceeds by the usual connective and quantifier rules. For instance, $\sim P\ulcorner\alpha\urcorner$ is to be true iff $P\ulcorner\alpha\urcorner$ is false, $\forall xPx$ is to be true iff $P\ulcorner\alpha\urcorner$ is true for every α, and so on. Provided that our language is first-order or first-order-like, these rules ought in principle to determine each sentence's *semantic value*, that is, how it stands with respect to truth and falsity.

Famously though, if the language contains a truth predicate T, all does not go according to plan. For equipped with T, one can form a sentence

(L) $\quad \sim T\ulcorner L\urcorner$

describing itself as untrue; and this sentence L (call it the *Liar sentence*) is highly resistant to semantic evaluation. The argument is in three steps. Because L and $\sim T\ulcorner L\urcorner$ are identical, their semantic values must be identical as well:

(1) $\quad \|L\| = \|\sim T\ulcorner L\urcorner\|.$

But the semantic value of a negation depends in a uniform way on its negatum's value, and likewise for a truth-sentence 'ϕ is true' and ϕ itself. Therefore the semantic value of 'ϕ is not true' is some function N of ϕ's semantic value; and this holds in particular when ϕ is the Liar:

(2) $\quad \|\sim T\ulcorner L\urcorner\| = N(\|L\|).$

Putting (1) and (2) together, and writing **s** for $\|L\|$, we get

(3) $\quad \mathbf{s} = N(\mathbf{s}).$

But in the context of classical semantics, where the semantic values are **t** and **f**, equation (3) is unsolvable.[7] So, *whatever* we do with the truth predicate in this case is going to violate some rule of evaluation.

Not so fast, one might say. Rather than calling (3) unsolvable, we could construe it as a challenge to the classical inventory of semantic values. This approach takes inspiration from the mathematician's response to another seemingly unsolvable equation: $\mathbf{x}^2 + 1 = 0$, or letting $n(\mathbf{x})$ be minus one divided by **x**,

(4) $\mathbf{x} = n(\mathbf{x})$.[8]

No *classical* (read *real*) number solves (4), so, the mathematician urges, let us look for a *non*classical number that does so. Such a number can of course be "found": $\mathbf{i} = \sqrt{-1}$. (And once you have **i**, you see that $-\mathbf{i}$ will also serve.)

Mightn't something similar work with (3)? Suppose that we fall in with the common assumption that a sentence's truth-sentence receives the same truth value(s) as the sentence itself:

(5) 'ϕ is true' should be counted $\left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\}$ iff ϕ is $\left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\}$.

Then as the following table makes clear, when ϕ is neither true *nor* false ($\overline{\mathbf{t}}\overline{\mathbf{f}}$), or both true *and* false (**t f**), 'ϕ is not true' shares these properties:

(6)

ϕ	$T\ulcorner\phi\urcorner$	$\sim T\ulcorner\phi\urcorner$
$\mathbf{t}\ \overline{\overline{\mathbf{f}}}$	$\mathbf{t}\ \overline{\overline{\mathbf{f}}}$	$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$
$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$	$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$	$\mathbf{t}\ \overline{\overline{\mathbf{f}}}$
$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$	$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$	$\overline{\overline{\mathbf{t}}}\ \mathbf{f}$
$\mathbf{t}\ \mathbf{f}$	$\mathbf{t}\ \mathbf{f}$	$\mathbf{t}\ \mathbf{f}$

This suggests that, instead of *two* semantical values, we allow *three* or perhaps *four*: uniquely true, uniquely false, neither true nor false (gap), and both true and false (glut). Then we can solve our equation $\mathbf{s} = N(\mathbf{s})$ by letting **s** be either of the two new values.[10]

III.

By the simple device of expanding the set of values, the Liar can be evaluated in complete conformity with our obligations—or so it seems, until

we realize that (5), hereafter the *straight* rule for truth, gets those obligations seriously wrong.

Take any sentence ϕ that is neither true nor false.[11] Then ϕ is at the very least *not true*; so to say that it *is* true should be to say something false. According to the straight rule, though, it is *not* false to call ϕ true. There is an opposite problem if some sentence ψ manages to be both true *and* false. Since ψ is among other things true, to *call* it true should *not* be to say anything false. But this is the evaluation that the straight rule insists on. As an ironic corollary, the "straight solution" to the paradox does not *itself* receive, under the straight rule, its intended semantic value. For remember, the straight rule instructs us to call *untrue* the description of a gap as untrue, or unfalse, and *false* the description of a glut as true, or false. So, if the Liar has neither truth value (both truth values), it will be untrue (false) to say so. Not only does the straight rule mistake our obligations, then, it ends up being in a good sense self-defeating.[12]

What would be a better statement of our obligations regarding the truth predicate? With any other predicate, the truth value of 'α is P' is *exclusively* a function of whether α is P. The straight rule breaks this pattern, letting 'α is true''s truth value depend not only on whether ϕ is true *but also on whether ϕ is false*. Nothing would seem to justify this special treatment except the fact that otherwise, we would get paradoxes. Since paradoxes are what we *do* get, this is no justification at all; therefore let us replace the straight rule (5) with the *strong* rule

(7) 'ϕ is true' should be counted $\left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\}$ iff ϕ is $\left\{\begin{matrix}\text{true}\\ \text{not true}\end{matrix}\right\}$,

and the straight truth table (6) with

(8)

ϕ	$T\ulcorner\phi\urcorner$	$\sim T\ulcorner\phi\urcorner$
t $\overline{\overline{\text{f}}}$	t $\overline{\overline{\text{f}}}$	$\overline{\overline{\text{t}}}$ f
$\overline{\overline{\text{t}}}$ f	$\overline{\overline{\text{t}}}$ f	t $\overline{\overline{\text{f}}}$
$\overline{\overline{\overline{\text{t}}\ \text{f}}}$	$\overline{\overline{\text{t}}}$ f	t $\overline{\overline{\text{f}}}$
t f	t $\overline{\overline{\text{f}}}$	$\overline{\overline{\text{t}}}$ f

Such a move has clear advantages, not least that when ϕ is neither true nor false, 'ϕ is neither true nor false'[13] now comes out *true* rather than gappy. Yet as a look at the truth table confirms, ϕ and 'ϕ is not true' can no longer be provided with the same semantic value. This means that the Liar sentence now resists *all*

attempts at semantic evaluation, whether as true, as false, as both or as neither.[14]

Over the next few sections we'll try to implement these ideas formally, starting with the "straight" theory (essentially Kripke's theory) and proceeding to the "strong" alternative just sketched, the one that depicts our truth concept as inconsistent. Because the straight rule treats ϕ and its truth sentence alike, the straight theory will have a broadly *constructive* character in which truth values are passed along sentence to sentence until the process ultimately saturates. When we switch to the strong rule, though, the truth values sentences *lack* become as important to subsequent assignments as the ones they possess. Since the information that ϕ *lacks* this or that truth value comes from "reflect[ion] on the...process"[15] by which truth values are distributed, the strong theory will have a *reflective* aspect whereby the fact that certain truth values are *not* going to be assigned becomes a basis for assigning others.

IV.

Begin with a first-order language $\mathcal{L}$ containing a distinguished predicate T for truth. Except for this one predicate, the language is to be conceived as completely understood; that is, we know the meaning of every predicate P other than T, and we know the meanings of the language's connectives and quantifiers. Both kinds of knowledge may be treated as knowledge of appropriate rules. On the one hand we have rules specifying for each predicate P which objects P is to be true (false) of in a given fact-situation or, what comes to the same given our assumption of a canonical name $\ulcorner\alpha\urcorner$ for every object α, rules specifying which atomic sentences $P\ulcorner\alpha\urcorner$ shall be true in a world and which shall be false there:[16]

(A1) $\|P\ulcorner\alpha\urcorner\| \approx \mathbf{t} \Leftrightarrow \alpha \in \mathbf{P^t}(w)$
(A2) $\|P\ulcorner\alpha\urcorner\| \approx \mathbf{f} \Leftrightarrow \alpha \in \mathbf{P^f}(w)$.

On the other hand we have rules saying how the truth values of complex sentences are to depend on the truth values of simpler ones. So, a negation is to be true (false) iff its negatum is false (true):

(N1) $\|\sim\phi\| \approx \mathbf{t} \Leftrightarrow \|\phi\| \approx \mathbf{f}$
(N2) $\|\sim\phi\| \approx \mathbf{f} \Leftrightarrow \|\phi\| \approx \mathbf{t}$.

Disjunctions are to be true (false) iff at least one disjunct is true (both disjuncts are false):

(D1) $\|\phi \vee \psi\| \approx \mathbf{t} \Leftrightarrow \|\phi\| \approx \mathbf{t}$ or $\|\psi\| \approx \mathbf{t}$
(D2) $\|\phi \vee \psi\| \approx \mathbf{f} \Leftrightarrow \|\approx\| \approx \mathbf{f}$ and $\|\psi\| \approx \mathbf{f}$.

And universal generalizations are to be true (false) iff all their instances are true (some are false):

(∀1) $\|\forall x\phi\| \approx \mathbf{t} \Leftrightarrow \forall \alpha \in w\ \|\phi(\ulcorner\alpha\urcorner)\| \approx \mathbf{t}$
(∀2) $\|\forall x\phi\| \approx \mathbf{f} \Leftrightarrow \exists \alpha \in w\ \|\phi(\ulcorner\alpha\urcorner)\| \approx \mathbf{f}$.[17]

That leaves atomic sentences containing the truth predicate. According to the *straight rule* these deserve the same truth values as their embedded φs:

(T1) $\|T\ulcorner\phi\urcorner\| \approx \mathbf{t} \Leftrightarrow \|\phi\| \approx \mathbf{t}$
(T2) $\|T\ulcorner\phi\urcorner\| \approx \mathbf{f} \Leftrightarrow \|\phi\| \approx \mathbf{f}$.

For brevity we refer to the conjunction of (A1) and (A2) as rule (A), the conjunction of (N1) and (N2) as (N), and so on; and where no confusion results we refer to the above rules en masse as the *straight rules*. Is it possible to evaluate sentences in accordance with the straight rules' requirements? Kripke's great discovery was that this always *is* possible.[18]

Think of the straight rules as a system of *equations* that we are attempting to solve. What would solve them is the right kind of *valuation*: a relation $\|\bullet\|$ between sentences and truth values such that, putting $\langle\phi,v\rangle \in \|\bullet\|$ for $\|\phi\| \approx v$ throughout, each of the above equations holds true. So, a valuation solves (N1) just in case it contains *both* of $\langle\phi,\mathbf{f}\rangle$ and $\langle\sim\phi,\mathbf{t}\rangle$ or *neither* of them. The goal is a valuation solving all of the rules simultaneously. (The lower-case Greek letter ρ will be used as a variable over valuations.)

To start with what we *know* how to get, call a valuation *factual* iff it solves all of the rules but (T), that is, it contains exactly the nonsemantical atomic attributions mandated by the nonsemantical facts, and it respects the connective and quantifier rules. Now let ρ′ solve (T) *relative to* ρ iff

(T′) ρ′ makes $T\ulcorner\phi\urcorner$ true (false) ⇔ ρ makes φ true (false).[19]

Then for each valuation ρ, there is a unique factual ρ′ solving (T) relative to ρ. Pretty clearly the operator taking ρ to ρ′ (call it *jump*) has all and only the solutions as its fixed points.[20] For on the one hand, a valuation identical to its own *jump* is factual and treats $T\ulcorner\phi\urcorner$ and φ alike—which is all there is to solving the rules. And conversely, if ρ is a solution, then ρ is a factual valuation solving (T), and so a factual valuation solving (T) relative to ρ. And this is all there is to being a fixed point of the jump operator.

Now all we need to show is that the jump operator *has* fixed points. By its definition *jump* is monotonic in the sense of preserving order relations among valuations (if $\rho \subseteq \rho'$, then $jump(\rho) \subseteq jump(\rho')$). Therefore if ρ is *sound* (included in its *jump*), the sequence

$$\rho(\alpha) = \begin{cases} \rho & \text{if } \alpha = 0 \\ jump(\rho(\alpha-1) & \text{if } a \neq 0 \end{cases}$$[21]

obtained by repeatedly applying *jump* to ρ will be increasing.[22] The sequence can't grow indefinitely (there are only so many sentences to evaluate), so eventually we reach a valuation identical to its own jump. This shows that fixed

points exist if sound valuations do. But the null valuation is sound automatically. So *jump* has fixed points and the rules are solvable.

That gives the essence of the straight approach, but there is a wrinkle that will gradually be assuming more importance. The semantic obligations laid down so far can be rolled up into a single "norm of equality":

(E) evaluate sentences in accordance with a solution.

But in general the rules have *many* solutions, not all of them equally plausible. Imagine that Socrates and Plato accuse each other of dishonesty, for instance, each calling the other's statement untrue. Then some solutions will make Socrates's accusation true and Plato's false, some will do the reverse, some will assign no truth value to either statement, and some will make both statements simultaneously true and false. That most of these alternatives seem clearly wrong suggests that we have not unlocked the rules' full deontic potential.

What we have in the equality norm is a *global* constraint on our evaluative behavior: no particular attributions are identified as incorrect, only certain ensembles of them.[23] But while part of proper semantic conduct has to do with the *pattern* of one's attributions, part concerns the attributions taken one by one. Therefore (E) needs to be supplemented with a *local* norm bearing separately on each proposed act of evaluation. And the obvious thought is that we should[24]

(G) evaluate ϕ as v iff the rules *dictate* that $\|\phi\| \approx$ v.

The *dictated*, or *grounded*, attributions are the ones that *must* be correct if all atomic attributions faithful to the nonsemantical facts are correct, and correctness is preserved by the connective, quantifier and truth rules.[25]

Taking stock, two norms govern our evaluative conduct. The *equality* norm directs us to evaluate sentences in accordance with a solution. And the *grounding* norm directs us to attribute truth (falsity) to a sentence just when such an attribution would be grounded. Is it possible for these norms to come into conflict? Just such a conflict will arise below, so let us see how it is avoided here. Say that a valuation is grounded iff all its members are. Then since *jump* preserves groundedness,[26] and the null set is grounded, the least fixed point is grounded as well. So the least fixed point, a.k.a. the smallest solution, contains *only* grounded attributions. But it can also be shown to contain *all* grounded attributions.[27] So *the grounded attributions comprise a solution*. This is a happy result, since otherwise there would be no way of doing our local duty without violating our global one. The bad news is that our luck is about to run out.

V.

Going back to Tarski at least, the leading intuition about "true" has been that a sentence's truth-sentence should be equivalent to the sentence itself. But

the intuition's undoubted appeal comes at the cost of a certain looseness about the kind of equivalence involved.

Tarski held that an adequate definition of "true" should yield all biconditionals $T\ulcorner\phi\urcorner \leftrightarrow \phi$ linking sentences with their truth-sentences. Nowadays though these biconditionals no longer strike us as essential to the concept. This is partly because, as Tarski stressed, unless special precautions are taken at least one of the biconditionals will be logically contradictory.[28] (Even granting that our truth-concept is somehow inconsistent, the idea that it forces literal contradictions on us seems implausibly crude.) Another problem is that, depending on one's semantics for biconditionals, $T\ulcorner\phi\urcorner \leftrightarrow \phi$ *cannot* be true unless ϕ has one or another of the classical truth values.[29] But then nothing as fancy as the Liar was needed to show that no predicate could behave as the truth predicate was supposed to: *any* non-bivalent sentence ('the King of France is bald', say) would have done. The point, of course, is that if a single non-bivalent sentence can prevent the biconditional standard from being met, then that standard is set far too high. Luckily, there is a more realistic standard close at hand. What the equivalence intuition *really* requires is, not that $T\ulcorner\phi\urcorner \leftrightarrow \phi$ should always come out true, but that $T\ulcorner\phi\urcorner$ should be true iff ϕ is true.

That leaves the question of when $T\ulcorner\phi\urcorner$ should come out *false*. Some see it as a corollary of the equivalence intuition that $T\ulcorner\phi\urcorner$ should be false just when ϕ is false. But why? There is no *general* reason why sentences alike in their truth-conditions must be alike in their falsity-conditions. And there is no reason specific to the case either. On the contrary, ϕ and $T\ulcorner\phi\urcorner$ seem to be false in *different* circumstances, witness their contrasting behavior under negation: when ϕ is untrue but not false, the negation of $T\ulcorner\phi\urcorner$ is true but the negation of ϕ is a gap. So the correct rule is that $T\ulcorner\phi\urcorner$ should be false just in case ϕ is not true. Recasting all of this in our official notation, the equivalence intuition is our old friend

(T1) $\|T\ulcorner\phi\urcorner\| \approx \mathbf{t} \Leftrightarrow \|\phi\| \approx \mathbf{t}$.

Thus the *truth*-conditions of truth-sentences remain the same. (Valuations that conform to (T1), and that are also factual, will be called *truthful*.) But the *falsity*-conditions of truth-sentences are modified from

(T2) $\|T\ulcorner\phi\urcorner\| \approx \mathbf{f} \Leftrightarrow \|\phi\| \approx \mathbf{f}$,

to

(T2) $\|T\ulcorner\phi\urcorner\| \approx \mathbf{f} \Leftrightarrow \|\phi\| \not\approx \mathbf{t}$.

Notice the change this entails in our project. Where before our goal was to find *factual* valuations satisfying (T1) and (T2), now it is *truthful* valuations satisfying (T2) that we are after; that is, truthful valuations making 'ϕ is true' false just when they fail to make ϕ true.

VI.

Usually in philosophy one is attempting something difficult but, the hope is, not impossible. Now we face a task that is both difficult *and* impossible—for not only are there problems about how to *search* for a solution to the present rules, there will not normally *be* any such solutions. These two aspects of our predicament are largely independent of each other. Even where solutions exist, there may be trouble locating them. And even where they do *not* exist, there can still be more or less reasonable ways of seeking them out, turning up more or less useful information en route. For the time being, I want to *bracket* the existence problem and to concentrate on devising workable search procedures.

How shall we bring the strong rule for truth to bear on the construction of valuations? To count 'ϕ is true' false on the basis that ϕ has not been found true *so far* would be risky, for if ϕ is *later* found true then the earlier attribution will be undermined. Should we wait for evidence that ϕ will *never* be found true? Such an approach seems hopeless. Whether ϕ will *later* be found true depends on what truth values we assign *now*; so our decision whether to count 'ϕ is true' false would require knowledge of the *outcome* of a process in which that decision itself figured. The result would be a vicious circle in which, to determine a sentence's truth value, we would need to have its truth value already in hand.

Ultimately there may be something to this charge of circularity. But there is a particularly striking *form* of the charge that can in fact be met. According to the rules for truth and negation, the truth of some sentences flows from the untruth of certain others: 'ϕ is not true' is true because ϕ is not true. From this it seems that one cannot identify every truth until one has first identified every untruth. On the other hand, it is not obvious how one can identify *any* untruths without first identifying every truth. (Until then, each candidate for untruth is potentially a not-yet-recognized truth.) But, if it takes all of the untruths to find all of the truths, and all of the truths to find *any* of the untruths, how will it be possible to locate even a single untruth?

To answer this we need a way of identifying untruths "in midstream," that is, *before* all of the truths are in. Here is my idea. Earlier we said that *two* norms govern the evaluation of sentences, the equality norm

(E) evaluate sentences in accordance with a solution,

and the grounding norm

(G) evaluate ϕ as v iff the rules dictate that $\|\phi\| \approx$ v.

That was before we revised the *straight* rules (A)-($\forall$), (T1) and (T2) to the *strong* ones (A)-($\forall$), (T1) and (T2)—but let's assume that *mutatis mutandis*, both norms still apply. Then we can use the second norm to solve the problem raised by the first, namely, how to detect untruths short of noting their absence from a completed list of truths. For if a case can be made that the strong rules *fail* to

dictate that $\|\phi\| \approx \mathbf{t}$, then the second norm tells us that ϕ is *not* true.

So far, so good, but now we need information about what the strong rules dictate. Because rules (A)-(∀) and (T1) are unproblematic, the question boils down to this: when do the strong rules dictate that a sentence ϕ's truth-sentence is false?

Imagine that we have somehow got a hold of a set of dictated attributions (the null set will do). By inspection of the rules, and especially rule (T2), they can never dictate both that $\|T\ulcorner\phi\urcorner\| \approx \mathbf{f}$ and that $\|\phi\| \approx \mathbf{t}$. But then the rules dictate that $\|T\ulcorner\phi\urcorner\| \approx \mathbf{f}$ only if our dictated attributions do *not* include an attribution of truth to ϕ. In this way a *lower* bound on the set of all dictated attributions gives rise to an *upper* bound on the dictated attributions of falsity to T-sentences; and this, once closed under the remaining rules (A)-(∀) and (T1), becomes an upper bound on the set of all dictated attributions whatsoever.

What the above shows is that an *under*estimate of the set of dictated attributions can be parlayed into an *over*estimate of that set. Mathematically the procedure harks back to section IV. There we defined the *jump* of a valuation as the smallest factual valuation solving (T) relative to the original.[30] That is, *jump*(ρ) =

> the least ρ' satisfying (A)-(∀) such that $\rho'(T\ulcorner\phi\urcorner) \approx \mathbf{v} \Leftrightarrow \rho(\phi) \approx \mathbf{v}$.

Now we let the *hop* of a valuation ρ be the smallest *truthful* valuation solving (T2) relative to ρ; that is, *hop*(ρ) =

> the least ρ' satisfying (A)-(∀) and (T1) such that $\rho'(T\ulcorner\phi\urcorner) \approx \mathbf{f} \Leftrightarrow \rho(\phi) \not\approx \mathbf{t}$.

Jump and *hop* are analogous in that, just as *jump*'s fixed points are the valuations solving the straight rules, *hop*'s fixed points are the valuations solving the strong rules. But notice two crucial differences between them, one technical and the other philosophical. First, where the jump operator *preserves* inclusion relations, the hop operator *reverses* them, turning the smaller valuation into the larger. (To have a word for this, the hop operator is *anti*monotonic.) Second, suppose that ρ contains only attributions dictated by the straight rules. Then *jump*(ρ) inherits this property.[31] But if ρ contains only attributions dictated by the strong rules, then *hop*(ρ), rather than containing *only* dictated attributions, contains *all* such attributions.[32] Accordingly every attribution outside of *hop*(ρ) is *not* dictated by the strong rules and therefore incorrect. *This solves our problem of how to identify untruths without first identifying all the truths.*

Now the hop operator, in addition to taking valuations containing *only* dictated attributions to ones containing *all* such attributions, takes the second kind of valuation back to the first. This allows us to set up two interlocking subprocesses, one approaching the set of dictated attributions from below, the other from above. Here is how to do it. Since *hop* is *anti*monotonic, the operator

skip obtained by composing *hop* with itself[33] will be *monotonic*. Define the lower and upper *skip*-sequences generated by a valuation ρ as follows:

$$\rho_\alpha = \begin{cases} \rho & \text{if } \alpha=0 \\ skip(\rho_{\alpha-1}) & \text{if } \alpha>0 \end{cases}$$

$$\rho^\alpha = \begin{cases} hop(\rho) & \text{if } \alpha=0 \\ skip(\rho^{\alpha-1}) & \text{if } \alpha>0.^{34} \end{cases}$$

For monotonicity reasons, if the initial valuation ρ is *sound* (included in its *skip*), the ρ_αs will gradually grow while the ρ^αs shrink; hence by the usual cardinality argument, we eventually reach a *lower* fixed point ρ_∞ and an *upper* one ρ^∞.[35] (These are called *duals* since each is the other's *hop*.) Assuming that the process *begins* from a set of dictated attributions, every attribution in the lower fixed point is dictated, and every dictated attribution is in the upper fixed point. Of course, the one starting point that *clearly* contains only dictated attributions is the empty set. All we can say for certain, then, is that attributions in the *smallest fixed point* are dictated, and attributions outside of the *largest fixed point* are not dictated.[36] (The first kind of attribution is called *definitely dictated* and the second *definitely not dictated*.)

Based on this construction, how close can we come to meeting our semantical obligations? There are four cases overall. First and unlikeliest, if the language is sufficiently impoverished then the least fixed point will be *identical* to the largest one. Equivalently the least fixed point is its own *hop*, which makes it a solution.[37] Since the least fixed point is *also* the set of dictated attributions,[38] this is a case in which we can evaluate sentences in accordance with a solution, as the equality norm demands, while still evaluating ϕ as v just when such an attribution is dictated, as required by the grounding norm.

Now suppose that the least fixed point is *distinct* from the greatest one—so that neither is a solution—but that a solution lies between them. This could happen, for instance, if the language was free of semantic anomaly except for the Double Liars:[39] the least fixed point would make *neither* Liar true, the largest fixed point would make *both* true, and the intervening solutions would make *one* of the two true at the expense of the other. The equality norm insists that we adopt one of these solutions, attributing truth to Socrates's utterance or, failing that, to Plato's. But since neither attribution is definitely dictated, this would be against the spirit, if not the letter, of the grounding norm.

Next, suppose that the least fixed point is distinct from the greatest one (so that neither is a solution); that there are no solutions between them; but that solutions do exist. This can be arranged by letting the language be free of anomaly except for the Truth Teller, which describes itself as true, and the Half Liar, which says that *either* the Half Liar is untrue or the Truth Teller is true.[40] *All* of the solutions on this scenario will evaluate the Truth Teller as true. Such an attribution is definitely *not* dictated,[41] though, so to do our global duty is in

this case to flout our local one.

The one remaining possibility is that the strong rules are *absolutely unsolvable*. (For example, the language might contain a Liar sentence.[42]) In this case our global duty cannot be carried out even at the expense of grounding.

Unless the smallest fixed point is identical to the largest one, it is not a solution. But it comes surprisingly close. Say that a valuation *makes* ϕ *true* (*false*) iff it contains $<\phi,\mathbf{t}>$ ($<\phi,\mathbf{f}>$); and *makes* ϕ *untrue* (*unfalse*) iff its dual omits $<\phi,\mathbf{t}>$ ($<\phi,\mathbf{f}>$). Then letting σ be the smallest fixed point,[43] we have

$$\sigma \text{ makes } P\ulcorner\alpha\urcorner \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \Leftrightarrow \alpha \text{ is in } \left\{\begin{matrix}\text{P's extension}\\ \text{P's antiextension}\end{matrix}\right\};$$

$$\sigma \text{ makes } \sim\phi \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \Leftrightarrow \sigma \text{ makes } \left\{\begin{matrix}\phi \text{ false}\\ \phi \text{ true}\end{matrix}\right\};$$

$$\sigma \text{ makes } \phi\&\psi \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \Leftrightarrow \sigma \text{ makes } \left\{\begin{matrix}\phi \text{ true and } \psi \text{ true}\\ \phi \text{ false or } \psi \text{ false}\end{matrix}\right\};$$

$$\sigma \text{ makes } \forall x\,\phi \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \Leftrightarrow \sigma \text{ makes } \left\{\begin{matrix}\text{each } \phi(\ulcorner\phi\urcorner) \text{ true}\\ \text{some } \phi(\ulcorner\phi\urcorner) \text{ false}\end{matrix}\right\};$$

$$\sigma \text{ makes } T\ulcorner\phi\urcorner \left\{\begin{matrix}\text{true}\\ \text{false}\end{matrix}\right\} \Leftrightarrow \sigma \text{ makes } \left\{\begin{matrix}\phi \text{ true}\\ \phi \text{ untrue}\end{matrix}\right\}.$$

This means that σ is in compliance with all of the rules *except* one direction of (T2), the one saying that a valuation should make 'ϕ is true' false if it does not make ϕ true. Instead σ has the slightly weaker property of making 'ϕ is true' false *iff it makes* ϕ *untrue*. Therefore our problem comes down to this. Sentences that are not made true by the least fixed point, or untrue either, do not have false truth-sentences as intuitively they should. The result is that 'ϕ is not true' does not come out true, although what it says is the case: ϕ is not among the true sentences.

VII.

Now that above ought to sound somewhat familiar. For near the end of his "Outline of A Theory of Truth," Kripke observes that his own theory runs into a similar problem:

> ...there are assertions we can make about the object language which we cannot make in the object language. For example, Liar sentences are *not true* in the object language, in the sense that the inductive process never makes them true; but we are precluded from saying this in the object language by our interpretation of negation and the truth predicate...[44].

The "interpretation" he refers to here is the straight rule for truth: 'ϕ is true' is true (false) iff ϕ is true (false). There is an "alternate intuition," however, which asserts that

> if ϕ is either false or undefined, then ϕ is *not true* and $T\ulcorner\phi\urcorner$ should be *false*, and its negation *true*...(80).

To accommodate this intuition, Kripke suggests a procedure called "closing off" the truth predicate:

> Take any fixed point... . Modify the interpretation of T(x) so as to make it false of any sentence outside [its extension]...(80)

The advantage of closing off is that "if ϕ is a paradoxical sentence, we can now assert $\sim T\ulcorner\phi\urcorner$." Yet if ϕ is the Liar paradox, then ϕ is *identical* to $\sim T\ulcorner\phi\urcorner$; so the sentence we are asserting to be untrue is the very one we are asserting — presumably as a truth. Kripke's explanation is that "the truth predicate of the closed off language defines truth for the fixed point *before* it was closed off." Of course, we can if we like "define a truth predicate for [the closed off] language in the usual Tarskian manner." But since the new predicate is like the old in describing (not the language in which it occurs but) a lower-level object language, we seem to be slipping into a Tarskian hierarchy of the sort we had thought to be done with. This is why Kripke says that "we still cannot avoid the need for a metalanguage."

Seen from our present perspective, Kripke's talk of an "alternate intuition" reflects a belated recognition of the strong rule's attractions. But the time for such a recognition is past; the straight rule has been in charge too long and no last minute salvage operation can undo its influence.[45]

Things look different, but not better, from the standpoint of a die-hard advocate of the straight rule. Under the straight rule, the die-hard points out, neither of our norms creates the slightest pressure for abandoning the least fixed point. After all, whoever patterns their conduct after the least fixed point is *already* evaluating sentences in accordance with a solution, and attributing truth (falsity) to a sentence just when the rules dictate this attribution.

Those adhering to the strong rule, however, *are* under pressure from the norms to venture beyond the least fixed point.[46] This is obvious in the case of the equality norm, since σ rarely succeeds in falsifying the truth-sentences of all untruths. But the grounding norm presses for action as well. Suppose for contradiction that σ contains *all* dictated attributions, and let ψ be a sentence that σ fails to make true, yet without making 'ψ is true' false. Since σ does not make ψ true, the rules do not dictate that ψ is true. By the grounding norm, therefore, ψ is *not* true, whence 'ψ is true' is false. Not only is this attribution correct, it is *dictated* on present assumptions.[47] Contradiction, since σ was supposed to *exhaust* the dictated attributions.

So we have something that the straight theorist does not: a rationale *from within our theory* for continuing beyond the least fixed point. But what sort of continuation is indicated? Start with the idea that there are *two* ways for a sentence to be untrue. Sometimes ψ *cannot* be true for reasons that emerge in the course of the construction: in this case σ *makes* ψ untrue, and 'ψ is true'

false as well. Sometimes though it is just that no reason ever emerges to evaluate ψ as true: in this case σ does not make ψ true, but since it does not make ψ untrue either, it does *not* make 'ψ is true' false. Preoccupied as we have been with the first kind of untruth, we have neglected the second, that is, *untruth by default*. But the norms observe no distinction here. *However* a sentence fails of truth, we are charged with evaluating 'ψ is true' as false, and with accepting the consequences under the remaining rules.[48] This whole operation[49] will be called *closing the fixed point off*. Basically it is the standard Liar reasoning writ large. For one is doing with *all* "default untruths" what is normally done with just the Liar: first, acknowledging that ψ is untrue; second, evaluating ψ's truth-sentence as false; and third, accepting the consequences, among them that 'ψ is not true' is true.

Now here is the wonder of the thing: technically, to close a fixed point off is just to apply the *hop* operator to it. This immediately tells us all we could want to know about the results. Closing the least fixed point off yields *another* fixed point (the largest one)[50] and this second fixed point meets (**T2**)'s requirement that the truth-sentences of all untruths ψ should be evaluated as false. On the minus side, as a result of falsifying these $T\ulcorner\psi\urcorner$s, the closed-off fixed point *verifies* some of the ψs that were previously assumed untrue. So we gain one direction of (**T2**) at the expense of the other. But the problems run deeper. When we classified all so-far-unverified sentences as untrue, that was because no case could be made for assigning truth to additional sentences. By so classifying them, though, we *created* a case for assigning truth to additional sentences. Worse yet, these new truths are *exactly* the sentences earlier deemed untrue by default![51]

Totaling up the damages, the assumption that the "default untruths" were indeed untrue has managed not only to destroy its own rationale, but directly to refute itself.[52] Doesn't this show that it was a mistake to begin with, and that it should now be withdrawn? *No* and *yes*, I claim. For reasons already indicated, we were *right* to count those sentences untrue. Having done so, though, and followed out the consequences, we have no option but to *reverse* ourselves and return to the original position[53]—whereupon our original rationale kicks back in and the cycle repeats, indefinitely. What drives this strange oscillation is, again, the strong rule for truth. Staying with the original fixed point—refusing to close it off—we would be admitting untruths with the incomprehensible property of being describable as *true* without risk of falsity. Refusing to undo the operation, we would be countenancing truths such that it was false to *call* them true. Either way we would be flouting our semantic duties, so we are forced to flip-flop.

Appendix 1

Notice something clumsy about our procedures: we have the language marching lockstep, as it were, through a series of wholesale reevaluations, but ordinary language evaluation is a much more flexible affair. This means, for example, that semantic values are assigned not all at once but sentence by sentence; and that one has choices about which semantical rules to apply and when. These appendices try to approximate the ordinary language situation by giving "deduction systems" for reasoning about sentences' semantic values.

Deduction systems are as usual collections of inference rules. But our notion of an inference rule is rather broader than the usual one. Traditional rules are sensitive only to what *is* accepted at a given dialectical juncture, never to what is *not* accepted. Naively, though, a premise's *unavailability* can be as inferentially relevant as the premise itself. ("Having no evidence to the contrary, I conclude that..."). This is especially so in the present case, where the fact that ϕ is *not* among the truths seems to license us in *inferring* that 'ϕ is untrue' is true. So our first departure from traditional logic is to allow *negative* inference rules, rules of the form:

> if such and such is accepted, but such and such else is *not* accepted, then draw such and such conclusions.

The second departure grows out of the first. Suppose my deduction system contains negative rules. Then it may happen that a conclusion is counted into my evolving theory *prematurely*, that is, on account of the absence of certain other statements, which, it turns out, are going to arrive later on. When these other statements make their appearance, my now baseless conclusion must be retracted, along with any further conclusions it may have sponsored. But how? Traditional rules are *additive* in the sense of permitting the derivation of new conclusions but never the *deletion* of old ones. Thus room must be made for *subtractive* rules:

> if such and such is accepted, draw such and such conclusions and *withdraw* such and such other conclusions.

So, a typical subtractive rule might instruct us to stop calling 'ϕ is true' false once we discover that ϕ is true. By a *nonmonotonic* rule we mean one that is negative or subtractive or both. Other rules, including the kind employed in standard logic, are *monotonic*.

To spell this out formally, start with an arbitrary language $\mathcal{K}$, to be identified with the set of its sentences. Quaternary relations $R(\Phi,\Psi,\Gamma,\Delta)$ among subsets of $\mathcal{K}$ are called *rules of inference*. Here is R's intended interpretation:

> if there are sets of sentences standing in R such that all of the Φs are accepted, and none of the Ψs are accepted, then add the Γs to the set of accepted sentences and remove the Δs from the set of accepted sentences.[54]

For readability, inference rules will be represented in the notation

$$\text{from}\begin{pmatrix}\Phi\\ \Psi\end{pmatrix},\ \text{infer}\begin{pmatrix}\Gamma\\ \Delta\end{pmatrix}\text{—where R }(\Phi,\Psi,\Gamma,\Delta);$$

except that if R is monotonic, so that Ψ and Δ are empty, we say simply: from Φ, infer Γ. By a *deduction system* Σ we mean an arbitrary collection of inference rules. Σ is *monotonic* (*positive, additive*) iff all its rules are monotonic (positive, additive).

Think of *deduction* as a process whereby one starts with a set of assumptions Θ_0, then transforms Θ_0 through sets Θ_1, Θ_2,... of interim hypotheses into some final collection Θ_γ of conclusions. This transformation process is guided by the deduction system Σ as follows. At any stage α, there is a set Θ_α comprising all currently accepted statements; its successor $\Theta_{\alpha+1}$ is chosen from among all sets obtainable from Θ_α via some applicable inference rule. Explicitly, a *deduction* D in system Σ is a sequence $<\Theta_\alpha|\alpha\leq\gamma>$, where

(a) Θ_0 = some set Ω;
(b) $\Theta_{\alpha+1}$ is obtained from Θ_α by some inference rule R, that is, for some $R \in \Sigma$ and Φ,Ψ,Γ,Δ standing in R, $\Phi \subseteq \Theta_\alpha$, $\Psi \subseteq \mathfrak{S}\Theta_\alpha$, and $\Theta_{\alpha+1} = (\Theta_\alpha \cup \Gamma) \sim \Delta$;[55]
(c) $\Theta_\gamma = \lim_{\alpha<\gamma} \Theta_\alpha$.[56]

Θ is *deducible* from Ω in Σ iff some Σ-deduction D begins with Ω and ends with Θ; and *provable* from Ω in Σ iff

(i) Θ is deducible from Ω, and
(ii) anything deducible from Θ has Θ as a subset.

So the sets provable from Ω are the deducible ones from which nothing can be removed by further application of the rules. Θ is deducible (provable) *simpliciter* iff it is deducible (provable) from the null set. Finally Θ is a *theory* of Σ iff it is maximal among sets provable in Σ.[57] Although I will not argue this here, positive systems always have at *most* one theory,[58] additive systems have in most cases at *least* one theory, and monotonic systems always have a *unique* theory.

Provability was just explained as a property of *sets* of sentences. When Σ is monotonic, there is a notion of proof for *individual* sentences as well. Say that a $\mathcal{K}$-sentence θ *follows from* a set Z of $\mathcal{K}$-sentences iff for some $R \in \Sigma$, $R(\Phi,\emptyset,\Gamma,\emptyset)$, $Z \subseteq \Phi$, and $\theta \in \Gamma$.[59] Tree is a *proof* of $\mathbb{T}$ in Σ iff[60]

(a) $\mathbb{T}$'s topmost node is occupied by θ;
(b) every sentence on $\mathbb{T}$ follows in Σ from the sentences directly below it; and
(c) $\mathbb{T}$ is grounded, in the sense of having no infinite branches.

Then it is a familiar fact that if Σ is monotonic, θ belongs to Σ's theory iff θ has a proof in Σ. This allows us to use 'θ is provable in Σ' indifferently for the existence of a proof of θ in Σ, and to claim membership for θ in Σ's unique theory.

Appendix 2

This appendix explains the principal fixed points of Kripke's *jump* operator, and of our own *skip* operator. After that, in the third and final appendix, we give deduction systems for these fixed points.

Already we have mentioned Kripke's *minimal* or *smallest* fixed point σ; this was the valuation generated by the null set under repeated application of *jump*. Among the sentences σ classifies as true are 'snow is white,' ''snow is white' is true,' '''snow is white' is true' is true'..., 'some statements are true,' and, should you deny the truth of any of the foregoing, 'some of *your* statements are untrue.' Among the sentences it classifies as gappy are: the Liar L (= 'L is not true'); the Double Liars L_s (= 'L_p is not true') and L_p (= 'L_s is not true'); the Truth Teller K (= 'K is true'); the Truth Waffler W (= 'K is true or K is not true'); and the Tautology Teller U (= 'U is true or U is untrue').

Now, since σ assigns the fewest truth values allowable under the rules, *all* fixed points find the σ-truths true. However not all fixed points share σ's views about the *second* group of sentences. Suppose we look at some of the options.

Opposite to σ lies a fixed point assigning as *many* truth values as the rules will allow. This *largest* fixed point τ makes the same sentences uniquely true (false) that σ does, while turning all of σ's gaps into gluts.[61] Since few have the heart to evaluate the Truth Teller (e.g.) as both true and false, the largest fixed point has been mostly ignored. Following Kripke, in fact, the practice has been to ignore all but *consistent* valuations, or valuations assigning at most one truth value to each sentence. The result is that instead of a largest fixed point we have a number of *maximal* ones χ_1, χ_2,...,[62] each internally consistent but inconsistent with all the rest.

When a sentence is true in some consistent fixed points but false in others, to attribute either truth value to that sentence would be arbitrary. The problem with maximal fixed points is that each of them is guilty of just such arbitrariness, by virtue of classifying the Truth Teller (e.g.) either as uniquely true or as uniquely false.

Then what about the *constant* attributions as a possible valuation, where an attribution is constant iff it occurs in some maximal fixed points and its opposite occurs in none?[63] These do not form a fixed point, unfortunately, since (e.g.) the Truth Waffler, a constantly true disjunction, lacks a constantly true disjunct.[64] But we are almost there. Say that an attribution is *intrinsic* iff it occurs in some fixed point *all* of whose members are constant. Then the intrinsic attributions *do* form a fixed point. This *largest intrinsic fixed point* ι agrees with σ in assigning no truth value to the Liar, the Double Liars, the Truth Teller, or the Truth Waffler. However the Tautology Teller is an intrinsic truth.

Surprisingly much of the above can be repeated, *mutatis mutandis*, in the context of strong theory. Where Kripke's smallest fixed point σ was obtained from the null set by repeated application of *jump*, our σ comes from the empty set by repeated application of *skip*. Provided that the extension of each predicate P other than T is disjoint from its antiextension, σ is included in σ: that is, *every truth value assigned by Kripke's minimal fixed point is assigned by ours*. Usually in fact σ assigns *more* truth-values than σ. For instance, if ϕ is a nonsemantic gap,[65] then $GAP\ulcorner\phi\urcorner =_{df} \sim T\ulcorner\phi\urcorner$ & $\sim T\ulcorner\sim\phi\urcorner$ is *gappy* in Kripke's minimal fixed point, but *true* in ours.[66] This is surely the right result: if ϕ is neither true nor false, it should be true to say so. More controversially, σ finds the Truth Teller K to be *false* rather than gappy.[67] This again seems to be the right result. Since 'I am true' is not, in and of itself, true, any sentence *calling* it true must be considered false. But K is itself such a sentence, so K is false.

Just now I said that σ assigns every truth value that σ does. This does not mean that σ assigns every *semantic* value that σ does, for not every σ-gap is a σ-gap. Sometimes the σ-gap receives a truth value, true or false, in σ; this is how it is with K, the Truth Teller. Other times σ leaves the σ-gap's semantic value undecided. This is how it is with the Liar sentence. The minimal fixed point does not make L true, but it does not make L *un*true either.

Keeping to the analogy with Kripke, let's confine ourselves to *coherent* fixed points, or fixed points making no sentence both true and untrue, or both false and unfalse.[68] Then opposite the smallest fixed point lie a number of *maximal* ones: χ_1, χ_2, etc. Except for a change of example necessitated by our reevaluation of the Truth Teller, the problem with the maximal fixed points is exactly what it was. Take the Double Liars L_s and L_p. For symmetry reasons these ought to be treated alike, but each χ_i supports one against the other, unbothered by the fact that other χ_js make the opposite choice.

Say that a sentence is *constantly* true (false) iff some maximal fixed points make it true (false) and none makes it untrue (unfalse). To insist on constant attributions is not sufficient since these do not form a fixed point—for example, the disjunction of the Double Liars is constantly true but lacks a constantly true disjunct.[69] But if we call those attributions *intrinsic* which belong to fixed points *all* of whose members are constant, then the intrinsic attributions do form a fixed point. As we would hope, this fixed point ι contains attributions that the least fixed point σ omits. Define the *moore sentence* of a sentence ϕ to be $\phi\&\sim T\ulcorner\phi\urcorner$, so that the moore sentence of 'it's raining out' is 'it's raining out but that's not true.' Then the sentence 'my moore sentence is untrue,'[70] though not a truth of the minimal fixed point, is true in the largest intrinsic fixed point.[71]

Appendix 3

This appendix gives deduction systems for reasoning about truth. To begin with our language $\mathcal{K}$ will be the set $\{\|\phi\| \approx v | \phi$ is a sentence of $\mathcal{L}$ and v is a truth-value$\}$ of *positive claims*. But later $\mathcal{K}$ will be expanded to include *negative claims* $\|\phi\| \neq v$ as well. These should be read as pronouncing it correct, or incorrect in the case of negative attributions, to assign the indicated truth values to the indicated sentences. For each set Θ of positive claims, and each fixed point ρ of *jump*, Θ *describes* ρ iff:

Θ contains the claim $\|\phi\| \approx \mathbf{t} \Leftrightarrow \rho$ makes ϕ true
Θ contains the claim $\|\phi\| \approx \mathbf{f} \Leftrightarrow \rho$ makes ϕ false

For each set Θ of positive and negative claims, and each fixed point ρ of the skip operator, Θ describes ρ iff along with the above we have

Θ contains the claim $\|\phi\| \not\approx \mathbf{t} \Leftrightarrow \rho$ makes ϕ untrue
Θ contains the claim $\|\phi\| \not\approx \mathbf{f} \Leftrightarrow \rho$ makes ϕ unfalse.[72]

The project is to find deduction systems whose theories describe the principal fixed points of the jump and skip operators.

Buried in a footnote of Kripke's "Outline of a Theory of Truth" is a system for calculating which attributions belong to σ = *jump*'s least fixed point.[73] This system, call it Σ_S, has the rules

[A1]	from ϕ,	infer $\|\mathbf{P}\ulcorner\alpha\urcorner\| \approx \mathbf{t}$ -- where $\alpha \in \mathbf{P}^{\mathbf{t}}(w)$
[A2]	from ϕ,	infer $\|\mathbf{P}\ulcorner\alpha\urcorner\| \approx \mathbf{f}$ -- where $\alpha \in \mathbf{P}^{\mathbf{f}}(w)$
[N1]	from $\|\phi\| \approx \mathbf{f}$,	infer $\|\sim\phi\| \approx \mathbf{t}$
[N2]	from $\|\phi\| \approx \mathbf{t}$,	infer $\|\sim\phi\| \approx \mathbf{f}$
[D1]	from $\|\phi_i\| \approx \mathbf{t}$ (i = 1 or 2),	infer $\|\phi_1 \vee \phi_2\| \approx \mathbf{t}$
[D2]	from $\|\phi_i\| \approx \mathbf{f}$ (i = 1 and 2),	infer $\|\phi_1 \vee \phi_2\| \approx \mathbf{f}$
[$\forall$1]	from $\|\phi(\ulcorner\alpha\urcorner)\| \approx \mathbf{t}$ (all α),	infer $\|\forall x\, \phi\| \approx \mathbf{t}$
[$\forall$2]	from $\|\phi(\ulcorner\alpha\urcorner)\| \approx \mathbf{f}$ (any α),	infer $\|\forall x\, \phi\| \approx \mathbf{f}$
[T1]	from $\|\phi\| \approx \mathbf{t}$,	infer $\|\mathbf{T}\ulcorner\phi\urcorner\| \approx \mathbf{t}$
[T2]	from $\|\phi\| \approx \mathbf{f}$,	infer $\|\mathbf{T}\ulcorner\phi\urcorner\| \approx \mathbf{f}$

Being monotonic, Σ_S has a unique theory Θ_S, which "is easily shown [to] correspond...to the minimal fixed point":[74]

(P1) Θ_S describes Kripke's minimal fixed point σ.

Recall from the end of the first appendix that a *proof* of $\|\phi\| \approx \mathbf{v}$ in Σ_S is a tree $\mathfrak{T}$ such that

(a) the claim that $\|\phi\| \approx \mathbf{v}$ occupies $\mathfrak{T}$'s topmost node,
(b) every claim on $\mathfrak{T}$ follows from the claims directly below it by some $R \in \Sigma_S$, and
(c) all of $\mathfrak{T}$'s branches are finitely long.

By the "fact" at the end of the first appendix, another way of formulating (P1) is this: it is provable in Σ_S that $\|\phi\| \approx \mathbf{t}$ (respectively, $\mathbf{f}$) iff the minimal fixed point makes ϕ true (false).

For the maximal fixed points a nonmonotonic system will be needed, that is, a system containing negative and/or subtractive rules. System Σ_X takes over the rules of Σ_S and adds the negative rule[75]

$$[X] \text{ from } \begin{pmatrix} \emptyset \\ \|\phi\| \approx \mathbf{v} \end{pmatrix}, \text{ infer } \begin{pmatrix} \|\phi\| \approx \bar{\mathbf{v}} \\ \emptyset \end{pmatrix}$$

and the subtractive rule

$$[Y] \text{ from } \begin{pmatrix} \|\phi\| \approx \mathbf{v} \\ \emptyset \end{pmatrix}, \text{ infer } \begin{pmatrix} \emptyset \\ \|\phi\| \approx \bar{\mathbf{v}} \end{pmatrix}$$

According to [X], $\|\phi\| \approx \bar{\mathbf{v}}$ can always be accepted, provided only that $\|\phi\| \approx \mathbf{v}$ is *not* accepted ("infer that $\|\phi\| \approx \bar{\mathbf{v}}$ whenever it is not immediately inconsistent to do so". But what if we find ourselves accepting $\|\phi\| \approx \mathbf{v}$ later on? This is where rule [Y] comes in: it allows us to abandon claims opposite to the claims that we accept. Using Θ_X as a variable ranging over Σ's theories:

(P2) the Θ_Xs describe the various maximal fixed points χ_1, χ_2, etc.

Deduction in Σ_X is messy but straightforward. Assuming for instance that we have

reasoned our way to Θ_S using the rules inherited from Σ_S, we might proceed by adding $\|L_s\| \approx \mathbf{t}$ via rule [X], then $\|L_p\| \approx \mathbf{t}$ by the same rule. But applying [N] and [T] to the former claim yields $\|\sim T\ulcorner L_s \urcorner\| \approx \mathbf{f}$, that is, $\|L_p\| \approx \mathbf{f}$. Now $\|L_p\| \approx \mathbf{t}$ is removable by rule [Y], and we are en route to a maximal fixed point backing Socrates's accusation against Plato's. By applying the rules in a different order, notice, we could have made Socrates out to be the liar.

Next comes Kripke's maximal intrinsic fixed point. The rules of Σ_I are [A]-[T] plus one additional rule, which requires explanation. By a *protoproof* of $\|\phi\| \approx \mathbf{v}$, let's mean a tree $\mathfrak{T}$ such that (a) $\mathfrak{T}$'s topmost node is occupied by $\|\phi\| \approx \mathbf{v}$, and (b) every claim on $\mathfrak{T}$ follows from the attributions directly beneath it by some $R \in \Sigma_S$. (So protoproofs are just like proofs, except that their paths are allowed to be infinitely long.) Protoproof $\mathfrak{T}$ is *consistent* iff no opposite claims $\|\phi\| \approx \mathbf{t}$ and $\|\phi\| \approx \mathbf{f}$ appear on it; and a consistent protoproof $\mathfrak{T}$ is *intrinsic* iff for every attribution $\|\phi\| \approx \mathbf{v}$ on $\mathfrak{T}$, no consistent protoproof of the opposite claim $\|\phi\| \approx \overline{\mathbf{v}}$ is possible. The new rule is this:

[I] from ϕ, infer $\|\phi\| \approx \mathbf{v}$ — where $\|\phi\| \approx \mathbf{v}$ has an intrinsic protoproof.

To see how [I] works, recall that the Tautology Teller U ($=T\ulcorner U \urcorner \vee \sim T\ulcorner U \urcorner$) is *intrinsically* true but not grounded true. Here is a protoproof of $\|U\| \approx \mathbf{t}$:

($\mathfrak{T}_1$) $\|U\| \approx \mathbf{t}$ (i.e., $\|T\ulcorner U \urcorner \vee \sim T\ulcorner U \urcorner\| \approx \mathbf{t}$) [D]

$\|T\ulcorner U \urcorner\| \approx \mathbf{t}$ [T]

$\|U\| \approx \mathbf{t}$ (i.e., $\|T\ulcorner U \urcorner \vee \sim T\ulcorner U \urcorner\| \approx \mathbf{t}$) [D]

$\|T\ulcorner U \urcorner\| \approx \mathbf{t}$ [T]

•
•
•

This is consistent, so it is intrinsic provided that attributions opposite to $\|U\| \approx \mathbf{t}$ and $\|T\ulcorner U \urcorner\| \approx \mathbf{t}$ do not also admit of consistent protoproofs. Imagine what a protoproof of $\|T\ulcorner U \urcorner\| \approx \mathbf{f}$ or $\|U\| \approx \mathbf{f}$ would look like:

($\mathfrak{T}_2$) $\|T\ulcorner U \urcorner\| \approx \mathbf{f}$ [T]

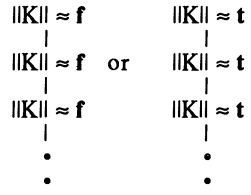
$\|U\| \approx \mathbf{f}$ [D]

$\|T\ulcorner U \urcorner\| \approx \mathbf{f}$ [T] $\quad \|\sim T\ulcorner U \urcorner\| \approx \mathbf{f}$ [N]

$\|U\| \approx \mathbf{f}$ [D] $\quad \|T\ulcorner U \urcorner\| \approx \mathbf{t}$ [T]

$\|T\ulcorner U \urcorner\| \approx \mathbf{f}$ [T] $\quad \|\sim T\ulcorner U \urcorner\| \approx \mathbf{f}$ [N] $\quad \|U\| \approx \mathbf{t}$ [D]

Since this is inconsistent, protoproof $\mathfrak{T}_1$ is intrinsic, whence by [I], the Tautology Teller is intrinsically true. Things are different with the Truth Waffler W ($= K \vee \sim K$). *Consistent* protoproofs of $\|W\| \approx \mathbf{t}$ are possible; but each contains a claim $\|K\| \approx \mathbf{v}$ whose opposite *also* admits of a consistent protoproof, viz.



Therefore [I] does *not* support the conclusion that W is true, which is what we would hope given that W is not intrinsic. More generally, ϕ is true (false) in Kripke's maximal intrinsic fixed point iff Σ's unique theory Θ_I pronounces it true (false):

(P3) Θ_I describes the maximal intrinsic fixed point ι.

This result would continue to hold if Σ_I consisted simply of rule [I]; the simplified system would be less convenient but it would prove exactly the same claims.

Now let's expand the language to include *negative* claims $||\phi|| \not\approx v$, intuitively to the effect that it is *in*correct to assign ϕ the truth value v. System ΣS takes over rules [A]-[$\forall$] and [T1] from ΣS, but [T2] is modified to

[T2] from $||\phi|| \not\approx t$, infer $||T\ulcorner\phi\urcorner|| \approx f$.

([T1] and [T2] together are called [T].) Note that [T2] will be useless unless *negative* claims are somehow provable. Define a *potential proof* of $||\phi|| \approx v$ as a tree $\mathbb{T}$ such that

(a) $\mathbb{T}$'s topmost node is occupied by $||\phi|| \approx v$;
(b) every positive claim on $\mathbb{T}$ follows from the claims directly below it by a single application of [A]-[$\forall$] or [T];
(c) all negative claims on $\mathbb{T}$ occupy terminal nodes; and
(d) all of $\mathbb{T}$'s paths are finitely long.

Say that a set Φ of positive claims *undermines* $||\phi|| \approx v$ iff each potential proof of $||\phi|| \approx v$ contains a claim in conflict with some Φ-claim. (Two claims *conflict* iff one is $||\psi|| \approx w$ and the other is $||\psi|| \not\approx w$.) Here is Σ_S's characteristic rule:

[U] from Φ, infer $||\phi|| \not\approx v$ -- where Φ undermines $||\phi|| \approx v$

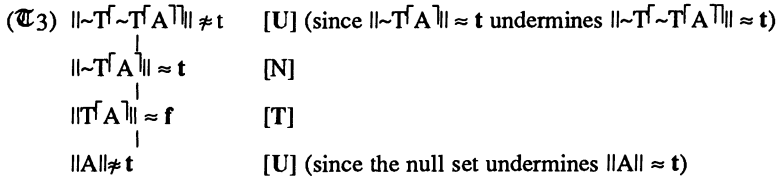
This completes the description of Σ_S. If Θ_S is its unique theory, then

(P4) Θ_S describes the minimal fixed point σ.

Recall that a *proof* of $||\phi|| \approx v$ ($||\phi|| \not\approx v$) in Σ_S is a tree $\mathbb{T}$ such that

(a) $\mathbb{T}$'s topmost node contains $||\phi|| \approx v$ ($||\phi|| \not\approx v$);
(b) every attribution on $\mathbb{T}$ follows from the attributions immediately below it by a single application of some $R \in \Sigma_S$; and
(c) all of $\mathbb{T}$'s paths are finitely long.

Another way of stating (P4) is that it is *provable* in Σ_S that $||\phi|| \approx t$ ($||\phi|| \approx f$, $||\phi|| \not\approx t$, $||\phi|| \not\approx f$) iff the least fixed point σ makes ϕ true (false, untrue, unfalse). Should A be a nonsemantic gap, for example, then we can prove as follows that σ makes $\sim T\ulcorner\sim T\ulcorner A\urcorner\urcorner$ untrue:



Substitute the Truth Teller K for A in this proof, and you get a proof that K is false,

that 'K is not true' is true, and that "K is not true' is not true' is not true.

Moving on to the maximal fixed points, our system for these takes the system Σ_S just given and adds two new rules. According to [X], whenever ϕ is not known to have truth value v, one may provisionally assume that it lacks that truth value. The point of [Y] is to enable us to withdraw that assumption, and any conclusions we may have drawn from it, should it turn out that ϕ is v after all:

$$\text{[X] from } \begin{pmatrix} \varnothing \\ \|\phi\| \approx v \end{pmatrix}, \text{ infer} \begin{pmatrix} \|\phi\| \not\approx v \\ \varnothing \end{pmatrix};$$

$$\text{[Y1] from } \begin{pmatrix} \|\phi\| \approx v \\ \varnothing \end{pmatrix}, \text{ infer} \begin{pmatrix} \varnothing \\ \|\phi\| \not\approx v \end{pmatrix};$$

$$\text{[Y2] from } \begin{pmatrix} \|\phi\| \not\approx v \\ \varnothing \end{pmatrix}, \text{ infer} \begin{pmatrix} \varnothing \\ \|\phi\| \approx v \end{pmatrix}.$$

Defining Σ_X as Σ_S plus [X] plus [Y], and using Θ_X as a variable over Σ_X's theories, we have

(P5) the Θ_Xs describe the various maximal fixed points χ_1, χ_2, etc.

Suppose for example that we have reasoned our way to Θ_S, describing the minimal fixed point. Since Θ_S omits $\|L_s\| \approx t$ and $\|L_p\| \approx t$, rule [X] lets us conclude both that $\|L_s\| \not\approx t$ and that $\|L_p\| \not\approx t$. The first claim yields $\|T^{\sqcap} L_s\| \approx f$ by [T] and thence $\|L_p\| \approx t$ by [N]; which by [Y] allows us to remove $\|L_p\| \not\approx t$. That done, there is no longer any possibility of removing $\|L_p\| \approx t$ or $\|L_s\| \not\approx t$, or of (re)introducing $\|L_p\| \not\approx t$ or $\|L_s\| \approx t$; so we are headed for a maximal fixed point endorsing Plato's claim against that of Socrates.

Only the maximal intrinsic fixed point remains. The rules of Σ_I are rules [A]-[$\forall$], [T], and [U] of Σ_S, and a new rule [I] which is explained as follows. A *protoproof* of $\|\phi\| \approx v$ ($\|\phi\| \not\approx v$) is a tree $\mathfrak{T}$ such that (a) $\mathfrak{T}$'s topmost node contains $\|\phi\| \approx v$ ($\|\phi\| \not\approx v$), and (b) every attribution on $\mathfrak{T}$ follows from those immediately below it by a single application of some $R \in \Sigma_S$. (So a protoproof is like a proof in system Σ_S except that it need not be grounded.) Protoproofs that do not contain conflicting claims $\|\psi\| \approx w$ and $\|\psi\| \not\approx w$ are called *coherent*; and a coherent protoproof is *intrinsic* iff for every claim it contains, no coherent protoproof of a conflicting claim is possible.[76] Rule [I] is this:

[I] from $\not\phi$, infer $\|\phi\| \approx v$ — where $\|\phi\| \approx v$ has an intrinsic protoproof.

Writing Θ_I for the system's one and only theory, we have

(P6) Θ_I describes the maximal intrinsic fixed point ι.

By (P1)-(P6), all of the principal fixed points can be built up piecemeal by application of appropriate inference rules.

To finish we sketch a deduction system Σ_T which, although it endorses exactly the attributions made by the minimal fixed point σ, invites us continually to reverse ourselves on the sentences σ finds impossible to classify—sentences (like the Liar) that σ does not make true, or untrue, or false, or unfalse. Σ_T starts with Σ_s and adds[77]

[V] from $\not\phi$, infer $\|\phi\| \not\approx v$ — where $\|\phi\| \approx v$ is not ΣS-provable.

This rule allows us to infer that ϕ is *not* true (false) whenever Σ_S does not prove that ϕ *is* true (false). Such a rule of course leads to incoherent conclusions: we can infer, for example, that the Liar sentence is not true, and then deduce by the other rules that it is true after all. Enter rule [Y] (see above), which permits us to abandon *either* of two opposing claims $\|\phi\| \approx v$, $\|\phi\| \not\approx v$ should we find ourselves at some point accepting both. Σ_T is Σ_S plus the two rules just stated.

Both systems have as their unique theory the same set of claims—a set that describes the least fixed point. The difference is that considerably *larger* sets, up to and including one describing the largest fixed point, are *deducible* in Σ_T, only to be

cut back down again by continued application of the rules. This same dialectic is exhibited in miniature by the standard Liar reasoning. Thinking of Θ as the set of claims accepted when the issue of how to evaluate L arises, we have:

(1)	$\Theta + \|\|L\|\| \neq \mathbf{t}$	(by [V])
(2)	$\Theta + \|\|L\|\| \neq \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f}$	(by [T])
(3)	$\Theta + \|\|L\|\| \neq \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f} + \|\|\sim T\ulcorner L \urcorner\|\| \approx \mathbf{t}$	(by [N])
(4)	$\Theta + \|\|L\|\| \approx \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f}$	(by [Y])
(5)	$\Theta + \|\|L\|\| \approx \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f} + \|\|T\ulcorner L \urcorner\|\| \neq \mathbf{f}$	(by [V])
(6)	$\Theta + \|\|L\|\| \approx \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \neq \mathbf{f}$	(by [Y])
(7)	$\Theta + \|\|L\|\| \approx \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \neq \mathbf{f} + \|\|L\|\| \neq \mathbf{t}$	(by [V])
(8)	$\Theta + \|\|L\|\| \neq \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \neq \mathbf{f}$	(by [Y])
(9)	$\Theta + \|\|L\|\| \neq \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \neq \mathbf{f} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f}$	(by [T])
(10)=(2)	$\Theta + \|\|L\|\| \neq \mathbf{t} + \|\|T\ulcorner L \urcorner\|\| \approx \mathbf{f}$	(by [Y])
(11)=(3)	.....	.....
.....	.....	.....;

and so on as far as desired. Such a pattern of to-ing and fro-ing is only to be expected of a sentence resisting all attempts at evaluation. As Wittgenstein suggests, this does not render the language any less usable. This is how the language is used.

Notes

1. "Agonist": a person who is torn by inner conflict, or (in physiology) a muscle whose action is opposed by that of another muscle. Ancestors of this paper were read in Saul Kripke's Truth Seminar at Princeton, and at Stanford's Center for Studies in Language and Information. Thanks to Nicholas Asher, Tyler Burge, John Etchemendy, Sally Haslanger, Saul Kripke, Tony Martin, Vann McGee, Leon Porter, and Jamie Tappenden.
2. This is roughly the *dialetheic* approach favored by Graham Priest (1979, 1984, 1987).
3. Although consider a dialetheic-sounding passage from *Lectures on the Foundations of Mathematics*: "I convince Rhees of the paradox of the Liar, and he says, "I lie, therefore I do not lie, therefore I lie and I do not lie, therefore we have a contradiction"" (§218). I find this reasoning implausible, because it forgets that Rhees is *reversing* himself at the second step: on concluding that he does not lie he *stops* thinking that he does lie.
4. *Zettel*, §686.
5. According to Gupta (1984, 1989), the meaning of "true" is constituted by a revision rule. I disagree. Revision occurs, and it occurs because of what "true" means. But, just as one can race around town *because* of one's commitments without being committed to racing around town as such, it is possible to *revise* because of the rules one is subject to without those rules' explicitly counseling revision. I say that the meaning of "true" is given by ordinary categorical rules, and that revision is a *derivative* phenomenon arising out of our attempts to do their bidding.
6. Or, better, aspects of its use, which depending on the type of word involved. So, the meaning of a predicate is a matter of the rules governing its application to objects, the meaning of a connective is given by rules for evaluating compound sentences based on their components' values, and so on. Because our topic is mainly predicates, I will be sloppy about the distinction between application-rules and rules of other kinds.
7. $N(\mathbf{t}) = \mathbf{f}$, and $N(\mathbf{f}) = \mathbf{t}$.
8. Spencer Brown (1973), pp. ix-x.
9. Here and throughout I interpret negation weakly: $\sim\phi$ is true (false) iff ϕ is false (true). (Note: $\underline{\underline{\mathbf{t}}}$ stands for untruth, $\underline{\mathbf{f}}$ for unfalsity.)
10. All possible positions are in fact occupied: Kripke says that the Liar is neither

true nor false, Graham Priest that it is both true and false, Peter Woodruff that it may be regarded either way (Kripke 1975, Priest 1979, Woodruff 1984). The situation is complicated by the fact that Kripke does not think of gap as a semantic value alongside truth and falsity. But all I mean by ϕ's semantic value is an answer to the questions: is ϕ true? is ϕ false? And when ϕ is the Liar, Kripke answers both questions in the negative. (Another view, which some have wanted to read into Kripke, is that the rules of evaluation leave both questions open. I find this view implausible but it is *not* the one I am arguing against in the text.)

11. For example, ϕ might be nonsense, or nonindicative, or guilty of presupposition failure.
12. Another, admittedly vaguer, worry is the following. Pretend for the moment that we use 'untrue' as (5) alleges. Still one can ask: is that the way it *had* to be, or might we equally have spoken according to (7) (see below)? This is important because if it is only by happenstance that the paradoxes are avoided, then although the immediate *semantical* challenge is met, *philosophically* we seem not further ahead. (Like remarks apply to approaches in which consistency is maintained through the intervention of some previously unnoticed semantical mechanism, for instance, approaches finding in 'true' an indexical or intensional element. See Burge (1979), Skyrms (1984), Barwise & Etchemendy (1987).)
13. Explicitly, $\sim T\ulcorner\phi\urcorner \,\&\, \sim T\ulcorner \sim\phi\urcorner$.
14. Suppose that I am wrong, and that 'ϕ is true' deserves the same truth-value(s) as ϕ. Still we *might* have used 'true' so that the truth-sentences of untruths were to be counted false, in which case the liar *would* have been the unsolvable paradox it appears to be. This makes the "straight solution" look somewhat uncurious. Rather than confronting the paradox in its natural habitat, the strong rule for truth, it seems content to emphasize our good fortune in being bound by the straight one. Of course, which rule is in fact operative is philosophically important. But the question remains: how can it happen that our semantical rules, though perfectly usable in the vast majority of cases, should give rise to paradoxes? To this, it is no help at all to say that we got lucky and they did not.
15. Kripke 1975, p. 80 (page references are to the version in Martin 1984).
16. '$\|\phi\| \approx \mathbf{v}$' means that ϕ deserves truth-value $\mathbf{v}$ (in w). The reason for using '$\approx$' rather than the identity sign is that we do not mean to rule it out that ϕ *also* deserves another truth-value (in that world). See Woodruff 1984. $\mathbf{P^t}(w)$ and $\mathbf{P^f}(w)$ are called P's extension and antiextension (in w).
17. Note that I am taking negation, disjunction, and universal quantification as basic; other connectives and quantifiers are to be regarded as defined from these in the usual manner. Also I use $\alpha \in w$ to mean that α exists in world w.
18. See also Martin & Woodruff 1975.
19. A valuation makes ψ true (false) iff it contains $\langle\psi,\mathbf{t}\rangle$ ($\langle\psi,\mathbf{f}\rangle$). The more usual way of putting (T′): ρ' *declares* true the sentences that ρ *makes* true.
20. An operator's fixed points are the objects $\mathbf{x}$ such that $O(\mathbf{x}) = \mathbf{x}$.
21. If α is β+1, $\rho(\alpha\text{-}1) = \rho(\beta)$. Otherwise $\rho(\alpha\text{-}1) = \cup_{\beta<\alpha} \rho(\beta)$.
22. "Increasing" means *nonstrictly* increasing: $\alpha \leq \beta \Rightarrow \rho(\alpha) \subseteq \rho(\beta)$.
23. Of course, particular attributions *are* indirectly affected, by way of the ensembles in which they figure. (Compare the EPA's Corporate Average Fuel Efficiency standards.)
24. Actually there are *two* obvious thoughts: first, as (G) says, all attributions not *required* by the rules should not be *permitted*; second, all attributions *permitted* by the rules should, as far as consistently possible, be *required*. The first thought leads to the least fixed point, the second to the maximal intrinsic fixed point. (See Appendix 2.)
25. That is, $\|\phi\| \approx \mathbf{v}$ is dictated iff (i) $\phi = P\ulcorner\alpha\urcorner$ and $\alpha \in \mathbf{P^v}(w)$, or (ii) $\|\phi\| \approx \mathbf{v}$ follows from type-(i) attributions by the remaining rules. Question: When do the rules dictate that $\|\phi\| \not\approx \mathbf{v}$? Answer: When they *don't* dictate that $\|\phi\| \approx \mathbf{v}$. This will be important later on.

26. Assume that ρ is grounded. Then each $\langle T\ulcorner\phi\urcorner, v\rangle$ such that $\langle\phi, v\rangle \in \rho$ is grounded; hence so is every attribution following from these $\langle T\ulcorner\phi\urcorner, v\rangle$s by (A)-($\forall$). The attributions so obtained make up *jump*(ρ).
27. *Every* solution has this property, just in virtue of being closed under the rules.
28. For example, the biconditional for the Liar sentence is $T\ulcorner L\urcorner \leftrightarrow \sim T\ulcorner L\urcorner$. Tarski's conclusion: "the concept of truth...when applied to colloquial language in conjunction with the normal laws of logic leads inevitably to confusions and contradictions" (1983, 165).
29. The remark applies to the weak and strong Kleene biconditionals. The Lukasciewicz biconditional can *in principle* be true when ϕ is neither true nor false; but it will *not* be true if I am right that ϕ's untruth makes $T\ulcorner\phi\urcorner$ false.
30. Actually we defined it as the *unique* factual valuation solving (T) with respect to the original; but since unique implies least, the definitions are equivalent.
31. See note 26.
32. This was argued in the last paragraph.
33. For all valuations ρ, *skip*(ρ) = *hop*(*hop*(ρ)).
34. I write $\rho^{\alpha-1}$ ($\rho_{\alpha-1}$) to mean: ρ^{β} (ρ_{β}), if $\alpha = \beta+1$, or $\cup_{\beta<\alpha}\rho^{\beta}$ ($\cup_{\beta<\alpha}\rho_{\beta}$), if α is a limit ordinal.
35. The words "lower" and "upper" are appropriate only if $\rho_{\infty} \subseteq \rho^{\infty}$. For this it suffices that ρ be *coherent* in the sense that it is included in its *hop*.
36. These are the fixed points generated by the null set.
37. Remember that the sets we are calling "fixed points" are the fixed points of *skip*; to be a solution you must be a fixed point of *hop*.
38. Note that since the smallest and largest fixed points are identical, every attribution is definitely dictated or definitely not dictated.
39. L_s describes L_p as untrue and vice versa.
40. The example is adapted from Gupta 1982.
41. As a matter of fact, the Truth Teller is *false* on the present theory (see Appendix 2).
42. Suppose that $\|\bullet\|$. is a solution. Then $\|L\| \approx \mathbf{t} \Leftrightarrow \|\sim T\ulcorner L\urcorner\| \approx \mathbf{t} \Leftrightarrow \|T\ulcorner L\urcorner\| \approx \mathbf{f} \Leftrightarrow \|L\| \not\approx \mathbf{t}$. Contradiction.
43. As a matter of fact, the stated equivalences hold for *any* fixed point.
44. This and succeeding quotes are from pp. 79-81 of Martin 1984, with inessential relettering.
45. Kripke informs me that one *can* reach my least fixed point from his, using a new operations he calls "almost closing off."
46. Here I mean the least fixed point not of *jump* but of *skip*.
47. Because the rules do not dictate that ψ is true, they *do* dictate that ψ is untrue (see note 25), and hence that 'ψ is true' is false.
48. That is, (A)-($\forall$) and (T1).
49. Here is the operation in more detail. First, gather together all the sentences not made true, including the "default untruths," and classify them as untrue. Second, apply (T2) to make the truth-sentences of these sentences false. Third, close under (A)-($\forall$) and (T1), the remaining rules.
50. Recall that the largest fixed point is τ = *hop*(σ).
51. *Proof*: ψ is untrue by default $\Rightarrow$ σ does not make ψ true or untrue $\Rightarrow$ $\langle\psi, \mathbf{t}\rangle$ is neither in σ nor outside $\tau \Rightarrow$ $\langle\psi, \mathbf{t}\rangle$ is in τ = the closed-off fixed point. Conversely, $\langle\psi, \mathbf{t}\rangle$ is in the closed-off fixed point but not in $\sigma \Rightarrow \langle\psi, \mathbf{t}\rangle$ is in τ but not $\sigma \Rightarrow \sigma$ does not make ψ true nor untrue $\Rightarrow$ ψ is untrue by default.
52. Stretching some terms from epistemology, we could say that the new valuation is both an *undercutting* and a *rebutting* defeater of the assumption that generated it, that is, the default assumption.
53. Technically this is just another application of the hop operator.
54. R is *positive* iff $R(\Phi,\Psi,\Gamma,\Delta) \Rightarrow \Psi = \emptyset$; otherwise *negative*. R is *additive* iff $R(\Phi,\Psi,\Gamma,\Delta) \Rightarrow \Delta = \emptyset$; otherwise *subtractive*.
55. $\overline{X}$ is X's complement in $\mathcal{K}$; Y~X is Y intersected with $\overline{X}$.
56. The precise nature of the limit operation is intentionally left open because our

results do not depend on it. All that matters is that $\liminf_{\alpha<\lambda}\Theta_\alpha \subseteq \lim_{\alpha<\lambda}\Theta_\alpha \subseteq \limsup_{\alpha<\lambda}\Theta_\alpha$. As usual, $\liminf_{\alpha<\lambda}\Theta_\alpha =_{df} \cup_{\beta<\lambda}\cap_{\beta<\alpha<\lambda}\Theta_\alpha$, and $\limsup_{\alpha\lambda} =_{df} \cap_{\beta<\lambda}\cup_{\beta<\alpha<\lambda}\Theta_\alpha$.

57. Yablo (ms) uses different, more demanding, notions of proof and theory.
58. *Sketch of proof:* Suppose for contradiction that Θ and Θ' are maximal provable, but neither is a subset of the other. $\Theta\cup\Theta'$ is deducible from Θ (Θ') by applying the rules used to deduce Θ' (resp. Θ) from the empty set in the very same order. Θ is maximal provable $\Rightarrow$ $\Theta\cup\Theta'$ is not provable $\Rightarrow$ some set Θ'' deducible from $\Theta\cup\Theta'$ lacks a member θ either of Θ or of Θ'. Contradiction since Θ and Θ' are provable and Θ'' is deducible from both.
59. Note that Ψ and Δ are the empty set because Σ is monotonic.
60. *Def.* A *tree* $\mathfrak{T}$ for $\mathcal{K}$ is, first, a set $|\mathfrak{T}|$ of nodes; second, a strict partial ordering $>$ on $|\mathfrak{T}|$ such that some node bears $>$ to every other node (there is a topmost node), and for all nodes m, $\{n|n \geq m\}$ is totally ordered by $>$ and finitely large (from the topmost node to any other there is a unique finite path); and third, an assignment to every node in $|\mathfrak{T}|$ of a $\mathcal{K}$-sentence said to *occupy* the node. $\mathfrak{T}$ is called *grounded* iff all of its branches are finitely long.
61. See Woodruff 1984. Where the *least* fixed point was the limit of the *jump*-sequence generated by the empty set, the *greatest* one is the limit of the *jump*-sequence generated by the set of everything.
62. I assume that there is more than one maximal fixed point.
63. Two attributions are *opposite* iff one is $\langle\phi,\mathbf{t}\rangle$ and the other is $\langle\phi,\mathbf{f}\rangle$.
64. Every maximal fixed point must decide, quite arbitrarily, which of W's disjuncts to evaluate as true.
65. So, ϕ might be $P^{\ulcorner}\alpha^{\urcorner}$, where α belongs neither to P's extension nor to its antiextension.
66. *Proof*: $\langle\phi,\mathbf{t}\rangle \notin hop(\phi) \Rightarrow \langle T^{\ulcorner}\phi^{\urcorner},\mathbf{f}\rangle \in hop(hop(\phi)) = skip(\phi) \Rightarrow \langle\sim T^{\ulcorner}\phi^{\urcorner},\mathbf{t}\rangle \in skip(\phi) \subseteq \sigma$. By a similar argument, $\langle\sim T^{\ulcorner}\phi^{\urcorner},\mathbf{t}\rangle \in \sigma$. So σ makes $GAP(\phi) = \sim T^{\ulcorner}\phi^{\urcorner}$ & $\sim T^{\ulcorner}\sim\phi^{\urcorner}$ true.
67. *Proof*: $\langle K,\mathbf{t}\rangle \notin hop(\phi) \Rightarrow \langle T^{\ulcorner}K^{\urcorner},\mathbf{f}\rangle \in hop(hop((\phi)) \subseteq \sigma \Rightarrow \langle K,\mathbf{f}\rangle \in \sigma$ since $K = T^{\ulcorner}K^{\urcorner}$. Note that K's "dual" $K' = \sim T^{\ulcorner}\sim K'^{\urcorner}$ ('I am not untrue') is *true* in our minimal fixed point.
68. Equivalently, ρ is coherent iff it is a subset of $hop(\rho)$.
69. For instance, $L_s \vee L_p$ is constantly true, but the same cannot be said of either disjunct.
70. Explicitly, $M = \sim T^{\ulcorner}M \,\&\, \sim T^{\ulcorner}M^{\urcorner}$.
71. The valuation making M true and assigning no other truth values is sound, so it generates a fixed point ρ under application of *skip*. This ρ can be shown to be intrinsic.
72. Recall that ρ makes ϕ untrue (unfalse) iff $hop(\rho)$ fails to contain $\langle\phi,\mathbf{t}\rangle$ ($\langle\phi,\mathbf{f}\rangle$).
73. Kripke (1975), note 24.
74. *ibid.*
75. **v** is **t** if **v** is **f**, and **f** if **v** is **t**.
76. Note that "protoproof" and "intrinsic" are defined differently here than they were above. Briefly, the new definition of "protoproof" replaces Σ_S with Σ_S, and the new definition of "intrinsic" has "coherent" where the old one had "consistent."
77. Strictly speaking, [V] makes [U] unnecessary, but we retain it for convenience.

References

Barwise, J. & Etchemendy, J. 1987: *The Liar: An Essay in Truth and Circularity*. Oxford: Oxford University Press.

Burge, T. 1979: "Semantic Paradox," *Journal of Philosophy* **76**, 169-98; reprinted in Martin 1984, 83-117.

Gupta, A. 1982: "Truth and Paradox," *Journal of Philosophical Logic* **11**, 1-60; reprinted in Martin 1984, 175-235.

Gupta, A. 1989: "Remarks on Definitions and the Concept of Truth," *Proceedings of*

the Aristotelian Society **89**, 227-46.
Gupta, A. & Belnap, N. 1993: *The Revision Theory of Truth*. MIT Press: Cambridge.
Herzberger, H. 1982: "Notes on Naive Semantics," *Journal of Philosophical Logic* **11**, 61-102; reprinted in Martin 1984, 133-174.
Kripke, S. 1975: "Outline of a Theory of Truth," *Journal of Philosophy* **72**, 690-716; reprinted in Martin 1984, 54-81.
Martin, R. & Woodruff, P. 1975: "On Representing 'True-in-L' in L," *Philosophia* **5**, 213-17; reprinted in Martin 1984, 47-51.
Martin, R. 1984: *Recent Essays on Truth and the Liar Paradox*. Oxford: Clarendon.
Priest, G. 1979: "The Logic of Paradox" *Journal of Philosophical Logic* **8**, 219-41.
Priest, G. 1984: "The Logic of Paradox Revisited," *Journal of Philosophical Logic* **13**, 153-79.
Priest, G. 1987: *In Contradiction: A Study of the Transconsistent*. Dordrecht: Boston.
Skyrms, B. 1984: "Intentional Aspects of Semantical Self-Reference," in Martin 1984, 118-131.
Spencer Brown, G. 1973: *Laws of Form*. Bantam: New York.
Tarski, A. 1983: "The Concept of Truth in Formalized Languages," in J.H. Woodger, ed. *Logic, Semantics, Metamathematics*. Indianapolis: Hackett, 152-278.
Wittgenstein, L. 1964: *Remarks on the Foundations of Mathematics*. Basil Blackwell: Oxford.
Wittgenstein, L. 1967: *Zettel*. Basil Blackwell: Oxford.
Wittgenstein, L. 1975: *Lectures on the Foundations of Mathematics*. University of Chicago Press: Chicago.
Woodruff, P. 1984: "Paradox, Truth, and Logic (I)," *Journal of Philosophical Logic* **13**, 213-52.
Yablo, S. 1985: "Truth and Reflection," *Journal of Philosophical Logic* **14**, 297-349.
Yablo, S. forthcoming: "Definitions, Consistent and Inconsistent," *Philosophical Studies*, Special Issue on Definition
Yablo, S. ms: "Notes on Nonmonotonic Logic".