# Kripke's Theory of Truth

(for the strong Kleene scheme)

Presented by Christopher Gauker
Version: July 26, 2006

Contemporary discussions of truth often make reference to Kripke's theory of truth (presented by Kripke only his 1975 *Journal of Philosophy* paper). What Kripke did was show how to interpret a language in such a way that it contains its own truth predicate. The fact that the language he describes contains its own truth predicate is proved by means of something called a fixed-point theorem.

I have written this document and posted it on the web because it is hard to find a presentation of Kripke's theory that will be understandable even to readers who have had only a first course in logic. It is my intention that this presentation will be understandable to all readers who have been exposed to at least the following: The languages of ordinary predicate logic. Recursive definitions of truth in a structure for such languages. (Structures are also called models or interpretations.) The basic concepts and notation of set theory, such as curly brackets, membership ("$\in$"), inclusion ("$\subseteq$"), ordered *n*-tuples, relations and functions.

Kripke's theory of truth builds on a three-valued interpretation of a language. (So sentences may be neither true nor false (N) as well as (T) or false (F).) Various three-valued valuation schemes may be used. Here we will consider only the strong Kleene scheme, which is the only one most people care about.

I can take no credit for this presentation. It is nothing more than a distillation from the more general presentation in Anil Gupta and Nuel Belnap's book, *The Revision Theory of Truth* (MIT Press, 1993) (who in turn acknowledge debts to Fitting and Visser). Gupta and Belnap's presentation deals with a broad range of valuation schemes simultaneously (including four-valued ones). This presentation merely boils theirs down to the point where it deals with only the strong Kleene valuation scheme. By thus narrowing our scope, we avoid many complications.

In every other presentation that I know of, something frustrating happens at the last minute. Just when the crucial fixed-point theorem is about to be achieved, the author appeals to some recondite fact of set theory, such as the fact that there is no 1-1 mapping of the ordinals into any set. The present presentation also employs such an assumption at the crucial point, but it is one that can be easily grasped on the basis of the definitions given here, namely, Zorn's Lemma.

Another clear presentation, in a very different style closer to Kripke's own, is that in Keith Simmons's, *Universality and the Liar* (Cambridge University Press, 1993), although Simmons's is one of those that contains a frustrating step at last minute (viz., the appeal to the Axiom of Replacement on p. 51). Simmons's method uses

transfinite induction over the ordinal numbers in place of the algebraic concepts of the present proof. (If you don't know what that means, it doesn't matter here.)

At the end, I will also state what I take to be the main reason for dissatisfaction with Kripke's theory of truth.

The strong Kleene valuation scheme:

Let *L* be a language with the usual connectives ($\neg$, $\vee$, $\wedge$, $\exists$, $\forall$) and in which sentences are defined in the usual way.

Let a *structure* M be a pair $\langle D, \Sigma \rangle$,

where D, called a *domain*, is a set of objects in the world,

and $\Sigma$, called an *assignment*, = $\langle \Sigma^0, \Sigma^+, \Sigma^- \rangle$, where

for all names (individual constants) n of *L*, $\Sigma^0(n) \in D$,

for all *n*-ary predicates R of *L*,

$\Sigma^+(R) \subseteq D^n$ (i.e., the set of *n*-tuples formed from members of D),

$\Sigma^-(R) \subseteq D^n$, and

$\Sigma^+(R) \cap \Sigma^-(R) = \varnothing$.

($\Sigma^+(R)$ is the *extension* of R, $\Sigma^-(R)$ is the *antiextension* of R, and the intersection of the extension and the antiextension of a predicate is always empty.)

$\sigma$, called a *variable assignment* (for a given structure), is a function from the variables of *L* into the domain D, i.e., such that $\sigma(v) \in D$.

The *empty variable assignment* is a variable assignment having the empty set as its domain (so that it assigns nothing to anything).

$\sigma[v|o]$ is a variable assignment just like $\sigma$ except that $\sigma(v) = o$.

Define: $\pi(t) = \begin{cases} \Sigma^0(t) \text{ if t is a name.} \\ \sigma(t) \text{ if t is a variable.} \end{cases}$

A *strong Kleene valuation* $Val_{M,\sigma}$ is a function from the set of formulas of *L* into {T, F, N}, and is defined relative to a structure M and variable assignment $\sigma$ as follows:

(1) Where R is an n-place predicate and $t_1$, $t_2$, …, $t_n$ are terms (variables or names),

$Val_{M,\sigma}(Rt_1t_2\ldots t_n) = T$ iff $\langle \pi(t_1), \ldots, \pi(t_n)\rangle \in \Sigma^+(R)$,

$Val_{M,\sigma}(Rt_1t_2\ldots t_n) = F$ iff $\langle \pi(t_1), \ldots, \pi(t_n)\rangle \in \Sigma^-(R)$,

$Val_{M,\sigma}(Rt_1t_2\ldots t_n) = N$ otherwise.

(2) Where P is a formula,

$Val_{M,\sigma}(\neg P) = T$ iff $Val_{M,\sigma}(P) = F$,

$Val_{M,\sigma}(\neg P) = F$ iff $Val_{M,\sigma}(P) = T$,

$Val_{M,\sigma}(\neg P) = N$ otherwise.

(3) Where P and Q are formulas,

$Val_{M,\sigma}((P \vee Q)) = T$ iff either $Val_{M,\sigma}(P) = T$ or $Val_{M,\sigma}(Q) = T$,

$Val_{M,\sigma}((P \vee Q)) = F$ iff both $Val_{M,\sigma}(P) = F$ and $Val_{M,\sigma}(Q) = F$,

$Val_{M,\sigma}((P \vee Q)) = N$ otherwise.

Similarly for the other 2-place connectives.

(4) Where P is a formula,

$Val_{M,\sigma}(\exists vP) = T$ iff for some $o \in D$, $Val_{M,\sigma[v/o]}(P) = T$,

$Val_{M,\sigma}(\exists vP) = F$ iff for every $o \in D$, $Val_{M,\sigma[v/o]}(P) = F$,

$Val_{M,\sigma}(\exists vP) = N$ otherwise.

Similarly for $\forall$.

A *strong Kleene valuation* of *sentences* (as opposed to arbitrary formulas) is a function $Val_M$ from sentences into {T, F, N} as follows:

Where $\sigma$ is the empty variable assignment, $Val_M(P) = T$ iff $Val_{M,\sigma}(P) = T$, $Val_M(P) = F$ iff $Val_{M,\sigma}(P) = F$, and $Val_M(P) = N$ otherwise.

Definition of partial order: ≤ is a *partial order* on a domain $\mathcal{U}$ iff ≤ is a relation on $\mathcal{U}$ such that:

(i) for all m ∈ $\mathcal{U}$, m ≤ m,

(ii) for all m, n ∈ $\mathcal{U}$, if m ≤ n and n ≤ m, then m = n, and

(iii) for all m, n, o ∈ $\mathcal{U}$, if m ≤ n and n ≤ o, then m ≤ o.

(We can represent a partial order as a set of ordered pairs. The interpretation of "≤" need have nothing to do with numbers.)

Concepts pertaining to partial orders:

Suppose ≤ is a partial order on a domain $\mathcal{U}$, and suppose $\mathcal{W} \subseteq \mathcal{U}$, and x ∈ $\mathcal{U}$. Then:

(i) x is an *upper bound* of $\mathcal{W}$ in $\mathcal{U}$ relative to ≤ iff, for all y ∈ $\mathcal{W}$, y ≤ x. (x may not be in $\mathcal{W}$).

(ii) x is a *least upper bound* of $\mathcal{W}$ in $\mathcal{U}$ relative to ≤ iff x is an upper bound of $\mathcal{W}$ and for all upper bounds y of $\mathcal{W}$ in $\mathcal{U}$, x ≤ y.

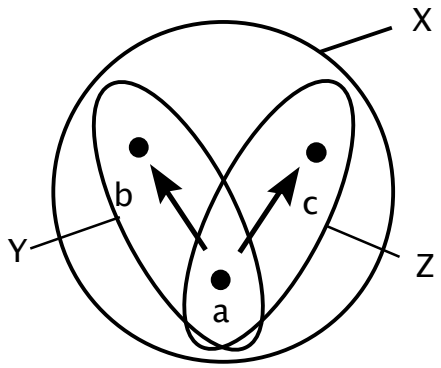
(iii) An element m ∈ $\mathcal{U}$ is a *maximal element* in $\mathcal{U}$ relative to ≤ if and only if there are no members n ∈ $\mathcal{U}$ such that n ≠ m and m ≤ n. (There may be no maximal element.)

(iv) $\mathcal{W}$ is a *chain* in $\mathcal{U}$ relative to ≤ if and only if for every x, y ∈ $\mathcal{W}$, either x ≤ y or y ≤ x.

If ≤ is a partial order on $\mathcal{U}$ and $\mathcal{W} \subseteq \mathcal{U}$, then $\mathcal{W}$ is *consistent* in $\mathcal{U}$ relative to ≤ iff for each two-membered set {m, n} ⊆ $\mathcal{W}$, {m, n} has an upper bound in $\mathcal{U}$.

≤ is a *coherent, complete partial order* (ccpo) on $\mathcal{U}$ iff $\mathcal{U}$ is partially ordered by ≤ and every consistent subset of $\mathcal{U}$ has a least upper bound in $\mathcal{U}$ relative to ≤.

Examples:




$\mathcal{X}$ = {a, b, c}, $\mathcal{Y}$ = {a, b}, $\mathcal{Z}$ = {a, c}

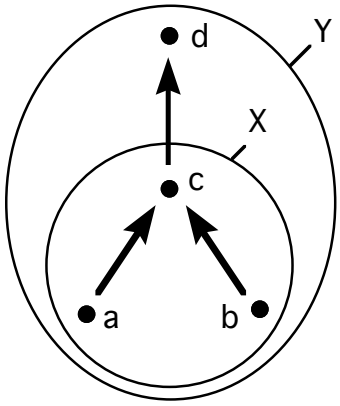
There is no upper bound of $\mathcal{X}$ in $\mathcal{X}$.

b is a least upper bound of $\mathcal{Y}$ in $\mathcal{Y}$ and in $\mathcal{X}$.

c is a least upper bound of $\mathcal{Z}$ in $\mathcal{Z}$ and in $\mathcal{X}$.

b is a maximal element of $\mathcal{Y}$, and c is a maximal element of $\mathcal{Z}$.

b and c are both maximal elements in $\mathcal{X}$.




$\mathcal{X}$ = {a, b, c}, $\mathcal{Y}$ = {a, b, c, d}.

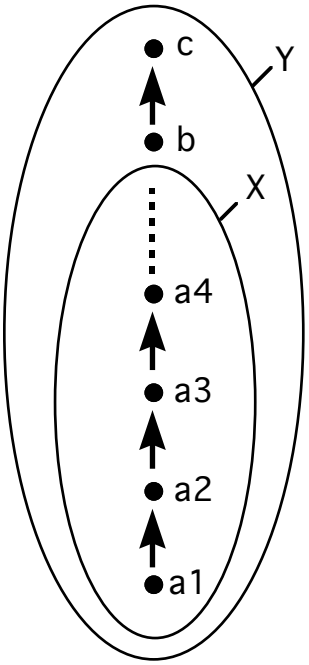
Both c and d are upper bounds of $\mathcal{X}$ in $\mathcal{Y}$.

c is a least upper bound of $\mathcal{X}$ both in $\mathcal{X}$ and in $\mathcal{Y}$.

d is a maximal element of $\mathcal{Y}$.

c is a maximal element of $\mathcal{X}$.

{a, c, d}, for example, is a chain in $\mathcal{Y}$, but {a, b, c} is not a chain in either $\mathcal{X}$ or $\mathcal{Y}$.




Suppose $\mathcal{X}$ = {a1, a2, a3, a4,…} is an infinite set, and both b and c are greater than every member of that set. $\mathcal{Y}$ = {a1, a2, a3, a4, …, b, c}.

b and c are both upper bounds of $\mathcal{X}$ in $\mathcal{Y}$, and b is the least upper bound of $\mathcal{X}$ in $\mathcal{Y}$.

There is no upper bound of $\mathcal{X}$ in $\mathcal{X}$, and there is no maximal element of $\mathcal{X}$.

c is a maximal element of $\mathcal{Y}$.

$\mathcal{X}$ is itself a chain (in $\mathcal{X}$ and in $\mathcal{Y}$), and $\mathcal{Y}$ is a chain in $\mathcal{Y}$.

$\mathcal{X}$ is consistent but not a ccpo (since $\mathcal{X}$ itself is a consistent subset of $\mathcal{X}$ but does not have a least upper bound in $\mathcal{X}$).

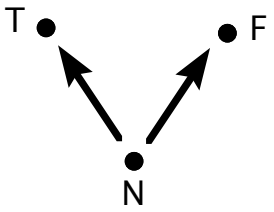
$\mathcal{Y}$ is consistent and also a ccpo.

Zorn's Lemma: Where ≤ is a partial order on $\mathcal{U}$, if every chain in $\mathcal{U}$ has a least upper bound in $\mathcal{U}$, then there are maximal elements in $\mathcal{U}$.

Note: Given the rest of the axioms of set theory, one can prove that Zorn's Lemma is equivalent to the Axiom of Choice (see Paul R. Halmos, *Naïve Set Theory*, Springer-Verlag, 1960, 1974, chapter 16), but one can also regard it as a basic, unprovable assumption.

The order of truth values

Let $\leq_{\mathcal{K}}$ be a partial order on the set of truth values $\mathcal{K}$ = {T, F, N} such that N is $\mathcal{K}$-less-than-or-equal-to both T and F and each value is $\mathcal{K}$-less-than-or-equal-to itself. That is, $\leq_{\mathcal{K}}$ = {⟨N, T⟩, ⟨N, F⟩, ⟨N, N⟩, ⟨T, T⟩, ⟨F, F⟩}. (So N $\leq_{\mathcal{K}}$ T and N $\leq_{\mathcal{K}}$ F, but also N $\leq_{\mathcal{K}}$ N, etc. F and T are not ordered with respect to one another.)




Interpreting the truth predicate

Suppose the language *L* contains the predicate "True".
Suppose we are given a structure M = ⟨D, Σ⟩ for a language *L* minus the predicate "True". (That is, take the sentences of *L*, remove all the sentences containing "True", and the result will be the language *L* minus "True".)
We stipulate that D contains all of the sentences of *L* as well as other things. (So sentences are themselves "objects in the world".)

Let $\mathcal{G}$ be the set of functions from D into {T, F, N}.

Let $\leq_{\mathcal{G}}$ be a relation on $\mathcal{G}$, where for all f, g ∈ $\mathcal{G}$, f $\leq_{\mathcal{G}}$ g iff for all o ∈ D, f(o) $\leq_{\mathcal{K}}$ g(o). (In other words, f is $\mathcal{G}$-less-than-or-equal-to g iff the truth value of o according to f is always $\mathcal{K}$-less-than-or-equal-to the truth value of o according to g. Of course, some pairs of members of $\mathcal{G}$ will not be ordered by this relation.) Obviously, $\leq_{\mathcal{G}}$ is a partial order on $\mathcal{G}$.

Where g is a member of $\mathcal{G}$, say that M+g is a structure $\langle D, \Sigma+g\rangle$,
where $\Sigma+g = \langle \Sigma^0, \Sigma^+ +g, \Sigma^- +g\rangle$, where
$\Sigma^+ +g$ is just like $\Sigma^+$ except that $\Sigma^+ +g(\text{"True"}) = \{\langle o\rangle \mid g(o) = T\}$, and
$\Sigma^- +g$ is just like $\Sigma^-$ except that $\Sigma^- +g(\text{"True"}) = \{\langle o\rangle \mid g(o) = F\}$.
In other words, M+g is a structure like M except that it also assigns an extension and antiextension to "True", which are determined by the function g. Note that it is not the function g that assigns an extension to "True" but the function $\Sigma^+ +g$.

We will take a special interest in one of these functions in $\mathcal{G}$, call it $g_0$, which is such that for all $o \in D$, $g_0(o) = N$.

We now define a function $\rho$ that takes us from one such function to another. Where g is a member of $\mathcal{G}$:

$$\rho(g)(o) = \begin{cases} \text{Val}_{M+g}(P), \text{ if } o = P, \text{ a sentence of } L. \\ F, \text{ if } o \text{ is not a sentence of } L. \end{cases}$$

In other words, for every nonsentence $o \in D$, $M+\rho(g)$ puts that object o into the antiextension of "True", since it is definitely not true. "That table is true" is thus false. For each sentence $o \in D$, $M+\rho(g)$ puts o into the extension of "True", the antiextension of "True", or neither, depending on whether o was true, false or neither in M+g. "$\rho$" is the Greek letter "rho". The function $\rho$ is called "jump", or "the jump function".

Note: It is not the case that every member of $\mathcal{G}$ other than $g_0$ is bound to be the result of applying $\rho$ to some other member of $\mathcal{G}$. In other words, the members of $\mathcal{G}$ do not form a chain ordered by $\rho$. One reason is that chains in $\mathcal{G}$ ordered by $\rho$ may have various starting points. For instance, in addition to the starting point $g_0$, we might have

a starting point that is just like $g_0$ except that T is assigned to the truth-teller, "This sentence is true". Another reason is that it can happen that there is a member of $\mathcal{G}$, g* such that for each of infinitely many members $g_i$ of $\mathcal{G}$, $g_i \leq_{\mathcal{G}} g^*$, even though g* is not the result of applying $\rho$ to any of the $g_i$. For instance, there could be a sentence that says, in effect, "The liar sentence is not true for any of the functions $g_0$, $\rho(g_0)$, $\rho(\rho(g_0))$, …", and that sentence might receive a truth value from g* even if it does not receive one from any of the functions $g_0$, $\rho(g_0)$, $\rho(\rho(g_0))$, … .

Definition:

A *fixed point* of a function f is an element e in the domain of f such that f(e) = e.

Our objective:

We want to show that $\rho$ has a fixed point. That is, there is a member of $\mathcal{G}$, call it $g_{\heartsuit}$ such that $g_{\heartsuit} = \rho(g_{\heartsuit})$.

What the significance of this result will be:

Call the structure $M+g_{\heartsuit}$ the *fixed-point interpretation* of *L*. Call a language interpreted by such a structure a *fixed-point language*. A fixed-point language contains its own truth predicate. That is to say, if P is a sentence in D and $\sigma$("x") = P, then $Val_{M+g_{\heartsuit},\sigma}(P) = Val_{M+g_{\heartsuit},\sigma}$("True(x)"). Consequently, if P is a sentence, then "True([" $\frown$ P $\frown$ "])" is true in *L* in $M+g_{\heartsuit}$ iff P is true in *L* in $M+g_{\heartsuit}$. So, contrary to what Tarski supposed (because he took bivalence for granted), a language can contain its own truth predicate.

Proof: Suppose $g_{\heartsuit}$ is a fixed point for $\rho$; suppose P is a sentence in D; and suppose $\sigma$("x") = P. Then, by the definition of $\rho$, $Val_{M+g_{\heartsuit},\sigma}(P) = T$ iff $\rho(g_{\heartsuit})(P) = T$. Since $g_{\heartsuit}$ is a fixed point for $\rho$, this is so iff $g_{\heartsuit}(P) = T$, which is so iff $\langle P \rangle \in \Sigma^+ + g_{\heartsuit}$("True"), which is so iff $Val_{M+g_{\heartsuit},\sigma}$("True(x)") = T. (Similarly, $Val_{M+g_{\heartsuit},\sigma}(P) = F$ iff $Val_{M+g_{\heartsuit},\sigma}$("True(x)") = F.)

Observation 1: $\leq_{\mathcal{K}}$ is a ccpo on $\mathcal{K}$ = {T, F, N}. Proof: Just check each consistent subset of $\mathcal{K}$ and see that it has a least upper bound.

Observation 2: For all g $\in \mathcal{G}$, $g_0 \leq_{\mathcal{G}} g$. This is so just because for all o $\in$ D, $g_0(o)$ = N and for all V in $\mathcal{K}$, N $\leq_{\mathcal{K}}$ V. In particular, $g_0 \leq_{\mathcal{G}} \rho(g_0)$.

Lemma 1: For all f, g $\in \mathcal{G}$, if $f \leq_{\mathcal{G}} g$, then for all sentences P of *L*, $Val_{M+f}(P) \leq_{\mathcal{K}} Val_{M+g}(P)$. (In other words, if g leaves no more gaps in the extension of "True" than f leaves, then no more sentences of *L* will be neither true nor false in M+g than were neither true nor false in M+f.) This can be proved by induction on the complexity of sentences.

Lemma 2: $\rho$ is monotone relative to $\leq_{\mathcal{G}}$.
That means: For all f, g in $\mathcal{G}$, if $f \leq_{\mathcal{G}} g$, then $\rho(f) \leq_{\mathcal{G}} \rho(g)$.
Proof: By Lemma 1, if $f \leq_{\mathcal{G}} g$, then for all sentences P of *L*, $Val_{M+f}(P) \leq_{\mathcal{K}} Val_{M+g}(P)$. But for all sentences P of *L*, $\rho(f)(P) = Val_{M+f}(P)$ and $\rho(g)(P) = Val_{M+g}(P)$. So for all sentences P of *L*, $\rho(f)(P) \leq_{\mathcal{K}} \rho(g)(P)$. For all nonsentences o $\in$ D, $\rho(f)(o) = \rho(g)(o)$ = F. So for all objects o $\in$ D, $\rho(f)(o) \leq_{\mathcal{K}} \rho(g)(o)$. So, $\rho(f) \leq_{\mathcal{G}} \rho(g)$.

Lemma 3: $\leq_{\mathcal{G}}$ is a ccpo on $\mathcal{G}$.
Proof: Suppose that $\mathcal{H}$ is a consistent subset of $\mathcal{G}$ (in the sense of "consistent" defined on p. 4 above). So if f, g $\in \mathcal{H}$, then {f, g} has an upper bound in $\mathcal{G}$. Since $\mathcal{H}$ is a set of functions from D into {T, F, N} and neither T nor F is $\mathcal{K}$-less-than-or-equal-to the other, this means that there is no object o $\in$ D such that either f(o) = T and g(o) = F or f(o) = F and g(o) = T.

We need to show that $\mathcal{H}$ has a least upper bound in $\mathcal{G}$ relative to $\leq_{\mathcal{G}}$. For each o $\in$ D, set $\mathcal{H}_o$ = {x| x = f(o) for some f $\in \mathcal{H}$}. By what I just explained about $\mathcal{H}$, $\mathcal{H}_o$ will

be either {T}, {F}, {N}, {T, N}, or {F, N}. By Observation 1, these all have a least upper bound in $\mathcal{K}$ relative to $\leq_{\mathcal{K}}$.

Define the function k as follows: For all o ∈ D, k(o) = the least upper bound of $\mathcal{H}_o$ (the existence of which we have just established). k will be a least upper bound for $\mathcal{H}$ in $\mathcal{G}$ relative to $\leq_{\mathcal{G}}$: By the definition of $\mathcal{G}$, k ∈ $\mathcal{G}$. k is an upper bound for $\mathcal{H}$: Since for every o ∈ D, k(o) is the least upper bound of $\mathcal{H}_o$, and for every other member f of $\mathcal{H}$, f(o) ∈ $\mathcal{H}_o$, for every o ∈ D, for every member f of $\mathcal{H}$, f(o) $\leq_{\mathcal{K}}$ k(o). Every other upper bound for $\mathcal{H}$ is "higher" than k relative to $\leq_{\mathcal{G}}$: Suppose that h is an upper bound for $\mathcal{H}$ in $\mathcal{G}$ and not(k $\leq_{\mathcal{G}}$ h). Then for some o ∈ D, not(k(o) $\leq_{\mathcal{G}}$ h(o)). Then since k(o) is the least upper bound for $\mathcal{H}_o$ relative to $\leq_{\mathcal{K}}$, h will not be an upper bound for $\mathcal{H}$.

Let $\mathcal{H}$ = {y | y is a function in $\mathcal{G}$ in the domain of ρ, and y $\leq_{\mathcal{G}}$ ρ(y)}.
In other words, something belongs to $\mathcal{H}$ just in case it is a function from D into {T, F, N} and it is $\mathcal{G}$-less-than-or-equal-to the function that results from applying jump to it.

Lemma 4: $\leq_{\mathcal{G}}$ is a ccpo on $\mathcal{H}$.
Proof: Let $\mathcal{J}$ be a consistent subset of $\mathcal{H}$. We need to show that $\mathcal{J}$ has a least upper bound in $\mathcal{H}$. $\mathcal{J}$ is a consistent subset of $\mathcal{G}$ as well. Since $\mathcal{G}$ is a ccpo, $\mathcal{J}$ has a least upper bound in $\mathcal{G}$. Call it b. To show that b is a least upper bound of $\mathcal{J}$ in $\mathcal{H}$, it suffices to show that b is a member of $\mathcal{H}$, i.e., that b $\leq_{\mathcal{G}}$ ρ(b). Let a be an arbitrary member of $\mathcal{J}$. Since b is an upper bound of $\mathcal{J}$ in $\mathcal{G}$, a $\leq_{\mathcal{G}}$ b. Since ρ is monotone (Lemma 2), ρ(a) $\leq_{\mathcal{G}}$ ρ(b). Since $\mathcal{J} \subseteq \mathcal{H}$, a $\leq_{\mathcal{G}}$ ρ(a). So a $\leq_{\mathcal{G}}$ ρ(b). So ρ(b) is an upper bound of $\mathcal{J}$ in $\mathcal{G}$. Since b is the least upper bound of $\mathcal{J}$ in $\mathcal{G}$, b $\leq_{\mathcal{G}}$ ρ(b).

Lemma 5: $\mathcal{H}$ has maximal elements.

Proof: Let $C$ be a chain in $\mathcal{H}$. By the definition of chains, $C$ is consistent in $\mathcal{H}$. Since $\mathcal{H}$ is a ccpo, $C$ has a least upper bound in $\mathcal{H}$. By Zorn's Lemma, $\mathcal{H}$ has maximal elements.

Kripke Fixed-Point Theorem: There are fixed points for $\rho$ in $\mathcal{H}$. In particular, there is a fixed point $g_{\blacktriangledown}$ for $\rho$ in $\mathcal{H}$ such that $g_0 \leq_{\mathcal{G}} g_{\blacktriangledown}$.

Proof: Let m be a maximal element in $\mathcal{H}$. By Observation 2 (since $\mathcal{H} \subseteq \mathcal{G}$), $g_0 \leq_{\mathcal{H}} m$. By Observation 2 again, $g_0 \leq_{\mathcal{G}} \rho(g_0)$. So $g_0 \in \mathcal{H}$. Since $m \in \mathcal{H}$, $m \leq_{\mathcal{G}} \rho(m)$. Since $\rho$ is monotone, $\rho(m) \leq_{\mathcal{H}} \rho(\rho(m))$. So $\rho(m) \in \mathcal{H}$. But m is maximal in $\mathcal{H}$. So $m = \rho(m)$. So m is a fixed point for $\rho$ such that $g_0 \leq_{\mathcal{G}} m$. Let $g_{\blacktriangledown} = m$.

The main criticism of Kripke's theory of truth

One of the main things one hopes for in a theory of truth is a diagnosis of the semantic paradoxes. Kripke's theory of truth takes us some distance toward that, but not very far. What Tarski showed (with his "undefinability theorem") is that a bivalent language cannot contain its own truth predicate. Consequently the liar sentence cannot be interpreted as saying what it seems to say, namely, that it itself does not belong to the extension of a predicate that subsumes all and only the true sentences of the same language to which the liar itself belongs.

What Kripke has shown is that if a language is not bivalent, then it can have a fixed-point interpretation on which it contains its own truth predicate. So the liar sentence can be interpreted as referring to itself and saying of itself that it belongs to the antiextension of a truth-predicate for very language to which it itself belongs. Further, we find, the liar sentence belongs to neither the extension nor the antiextension of the truth-predicate on such a fixed-point interpretation. Similarly, we can show that if a language is not bivalent, then it can contain both its own truth predicate and its own falsehood predicate. That is, there can be a structure and a predicate "is true" and a

predicate "is false" such that a sentence belongs to the extension of "is true" in that structure if and only if it is true in that structure and belongs to the extension of "is false" in that structure if and only if it is false in that structure.

The trouble is that such fixed-point languages still cannot contain theiry own non-truth predicates (as Kripke himself noted in his 1975 paper). That is, for every structure for such a language, there is no predicate in the language (for example, "is not true") such that a sentence belongs to the extension of that predicate in that structure if and only if the sentence is not true in that structure. I will prove that presently. But first, how can that be so if "is true" is a truth-predicate for a fixed-point language? Well, if "is true" is the truth predicate in a fixed-point language, then the extension of "is not true" comprises the objects in the antiextension of "is true", which unfortunately may not include all of the objects that fall outside of the extension of "is true".

The consequence is that Kripke's theory of truth really does not give us a diagnosis of the semantic paradoxes. If we think of the liar sentence as a sentence in a Kripke-fixed point language, then the predicate "not true" that occurs in it cannot be interpreted as meaning what we interpret it as meaning when we think of the liar sentence as belonging to *our own* language, namely, that the object of which it is predicated falls outside the extension of "true". There have been attempts to excuse this result (for example, in Scott Soames's book *Understanding Truth* (Oxford, 1998, pp. 188-190)), but in my opinion we do not need to make excuses because there is a better theory of truth to be had. (See my "Semantics for Deflationists", in JC Beall and Bradley Armour-Garb, *Deflationism and Paradox*, (Oxford University Press, 2005), pp. 148-176.)

I will now prove that Kripke's fixed point languages cannot contain their own non-truth predicate. More precisely, they cannot do that if they meet certain other minimal conditions that we should expect any language to meet if our semantics for that language is to provide a model (in the sense of paradigm case) for our diagnosis of the paradoxes that arise in our own language.

I will assume that the language in question contains a quotation name for every formula in that language. For clarity, however, I will use square brackets to form quotation-names rather than quotation marks. Thus, the name of "*x* is a tree" is "[*x* is a tree]", and, in general, if *S* is any formula of the language, then [*S*] is its quotation-

name. I also assume that the language contains its own *diagonal predicate*. If $F(v)$ is a formula of the language containing $v$ as its sole free variable and $[F(v)]$ is its quotation-name, then $F([F(v)])$ is the *diagonal* of $F(v)$. Let the diagonal predicate of the language be **D**, so that a sentence of the form **D***ab* means that the sentence that *a* denotes is the diagonal of the sentence that *b* denotes.

A well-known observation of Gödel's, sometimes called the diagonal lemma, tells us that, under these conditions, for any formula $F(v)$ of our language containing $v$ as its sole free variable, we can construct a *Gödel-sentence* $A$ for $F(v)$ such that $A$ is true if and only if $F([A])$ is true. In particular, the following sentence is such a Gödel-sentence for $F(v)$:

$\exists y(\mathbf{D}y[\exists y(\mathbf{D}yx \wedge F(y))] \wedge F(y))$.

It is evident that this sentence will be true if and only if the following sentence is true:

$F([\exists y(\mathbf{D}y[\exists y(\mathbf{D}yx \wedge F(y))] \wedge F(y))])$.

By virtue of the meaning of **D**, the first of these two sentences is true if and only if $\exists y(y = [\exists y(\mathbf{D}y[\exists y(\mathbf{D}yx \wedge F(y))] \wedge F(y))] \wedge F(y))$ is true, which is so if and only if the second sentence is true.

Suppose, for a reductio, that a Kripke fixed-point language contains its own non-truth predicate **NT**. Since **NT** is a non-truth predicate, we have it that:

(i) **NT**[*s*] is true in *L* if and only if *s* is either false in *L* or neither true nor false in *L*.

By the Gödel diagonal lemma, there is a sentence *A* of *L* such that

(ii) *A* is true in *L* if and only if **NT**[*A*] is true in *L*.

From (i) and (ii), we derive:

(iii) *A* is true in *L* if and only if *A* is either false in *L* or neither true nor false in *L*.

But (iii) is a contradiction. So we were mistaken to suppose that a Kripke fixed-point language might contain its own non-truth predicate.