Notre Dame Journal of Formal Logic
Volume 31, Number 2, Spring 1990

Logics of Truth

RAYMOND TURNER

Abstract This paper surveys three recent semantic theories of truth and compares them from the perspective of their underlying logics. In particular, the underlying logic of the Gupta–Herzberger theory is investigated, and an analysis of modal logics of truth arising from this semantic theory is given.

1 Theories of truth In recent years there has been a revival in the development of semantic theories of truth. They are all attempts to develop theories of truth for languages which contain their own truth predicates and moreover they are all semantic theories in that they are grounded in some semantic interpretation of the truth predicate. In this paper we shall compare these theories from the perspective of their underlying logics of truth. We shall concentrate on those theories cast within the framework of classical logic, since this is where the notion of truth is most at home. Three of the most influential theories are those of Scott [10]–Aczel [1], Kripke [9]–Gilmore [4]–Feferman [3], and Gupta [5]–Herzberger [6]. In the case of the first two kinds of semantic theories the logic of the system is explicit. The main objective of this paper is to explore the underlying logic of the last kind of theory and to explore its connections with the other two.

1.1 Frege structures Aczel [1] introduces the notion of a Frege structure in an effort to capture the consistent subtheory of Frege's *Grundgesetze der Arithmetik*. Aczel formulates his theory within a model of the untyped Lambda Calculus and develops a theory of classes or types based upon a theory of truth and propositions. Frege structures are models of the Lambda Calculus enriched with two subsets: a set of propositions and a set of true propositions. These sets satisfy very natural closure conditions with respect to the logical connectives: on

Received April 13, 1988; revised October 18, 1988

the class of propositions the truth predicate obeys the Tarski criteria. We shall briefly review the theory of Frege structures and use it as a way into the theory of Kripke–Gilmore–Feferman.

1.2 The Kripke–Feferman–Gilmore theory Various approaches to the semantic paradoxes result in some logical schema to capture the safe instances of the Tarski biconditionals. For example, the approach of Gilmore [4]–Feferman [3] has as consequences the schemas:

$$\mathrm{T}(A) \text{ iff } A^{+}$$
$$\mathrm{T}(\sim A) \text{ iff } A^{-}$$

where A^{+} and A^{-} are in some sense *approximations* to A and $\sim A$ respectively. This particular approach employs an inductive construction in the development of models for the theory. Moreover, even though the theory is cast within the general setting of classical logic the construction of the models is essentially based on the inductive technique introduced into truth-theory by Kripke [9], and gains its formal credibility through a nonstandard treatment of negation in wff's such as A^{+} and A^{-}. In essence these approximations are the result of pushing all negations into atomic position and replacing all such negations by *internal negations.* As a consequence the *internal* logic of the truth predicate is nonstandard. We develop a version of this theory in which its three-valued nature is made explicit.

1.3 The Gupta–Herzberger theory In contrast, the approach of Gupta [5] and Herzberger [6] is totally classical and employs a *semi-inductive* technique in the construction of the models. The Gupta–Herzberger theory of truth is based upon a notion of truth as revision. Herzberger refers to his approach as "naive semantics", in reference to those naive beliefs about the concept of truth which lead to the paradoxes. The theory of truth advocated is a modification of Kripke's theory within a classical framework. The theory thus employs only classical models and ordinary two-valued valuations. In Turner [11] a development of the logic underlying the Gupta–Herzberger theory was begun. There we alluded to the modal logic implicit in the semi-inductive process and even stated a version of the logic. The main aim of the present paper is to more fully investigate the logic of truth which underlies the Gupta–Herzberger semantics. We develop within the context of the Lambda Calculus the theory of truth developed in Gupta [5]–Herzberger [6]. We then investigate the various logics of truth which are sound under the Gupta–Herzberger semantic theory. The result is a family of modal logics of truth.

2 Truth and the Lambda Calculus In this section we introduce the formal background to the logics we later develop. We are primarily concerned with the development of a theory of truth for a language which contains enough formal machinery to represent the various versions of the Liar sentence and related paradoxical sentences. Perhaps the most elegant way to achieve this is to admit the Lambda Calculus as the language of terms.

2.1 The Lambda Calculus To begin, we present the basic background material on the Lambda Calculus, beginning with the language.

Basic Vocabulary

Individual variables $x, y, z, \ldots$
Individual constants $c, d, e, \ldots$

Inductive Definition of Terms

(i) Every variable or constant is a term
(ii) If t is a term and x is a variable then $(\lambda x.t)$ is a term
(iii) If t and t' are terms then (tt') is a term.

We adopt the following standard axiomatization of the β-Lambda Calculus, where $t[s/x]$ is the result of substituting s for every free occurrence of x in t.

Axioms of the $\lambda\beta$-Calculus

$\lambda x.t = \lambda y.t[y/x]$, y not free in t
$(\lambda x.t)t' = t[t'/x]$.

We require a few basic facts about the Lambda Calculus. We shall be brief and refer the reader to Hindley and Seldin [7] and Barendregt [2] for more details. We assume some standard representation of the pairing and projection combinators $\langle .,. \rangle$, fst, and snd which satisfy: $\text{fst}(\langle x, y \rangle) = x$ and $\text{snd}(\langle x, y \rangle) = y$, and some standard representation of the numerals (e.g., the Church representation) 0,1,2,3,4,5,.... We shall also appeal to the fixed point theorem of the Lambda Calculus.

Theorem 2.1.1 *There is a lambda term Y such that, for every lambda term t,* $t(Yt) = Yt$.

As regards the models of the Lambda Calculus, for the sake of concreteness we employ a version of Scott models. The central notion is the following:

Definition 2.1.2 A *domain* is a partially ordered set, with a least element u, which admits the least upper bounds of ω-sequences.

We can spell this out in a little more detail. Let D be a domain and let $\subseteq$ be the ordering of the domain. The element u is the *least element* of the domain D if $u \subseteq d$ for each d in D. An ω-sequence is of the form $d_0 \subseteq d_1 \subseteq \ldots \subseteq d_n \subseteq \ldots$, where $d_i \in D$ for $i \geq 0$. An element d is an *upper bound* of the sequence if $d_i \subseteq d$ for each $i \geq 0$; it is a *least upper bound* if $d \subseteq d'$ for any other upper bound d' of the sequence. We write the least upper bound of the sequence as $\bigcup_i d_i$.

Definition 2.1.3 A function $f: D \to D'$ is *continuous* iff for each ω-sequence $\langle d_n \rangle_{n \in \omega}$ in D, $f(\bigcup_i d_i) = \bigcup_i f(d_i)$.

The next stage in the construction is to indicate how the class of continuous functions from D to D' (where D and D' are arbitrary domains) forms a domain.

Definition 2.1.4 Let $[D \to D']$ be the class of continuous functions from D to D'. For $f, g \in [D \to D']$ define:

$$f \subseteq g \leftrightarrow (\forall d \in D)(f(d) \subseteq' g(d)),$$

where $\subseteq'$ is the ordering of D'.

Theorem 2.1.5 *$[D \to D']$ with the above ordering forms a domain.*

This completes the basic preliminaries. In the present context a *Scott model* is a domain D which is isomorphic to its own continuous function space where the isomorphisms are themselves continuous. More precisely:

Definition 2.1.6 A *Scott model* is a triple $\boldsymbol{D} = \langle D, \Phi, \Psi \rangle$ where
(i) D is a domain
(ii) $\Phi : D \to [D \to D]$ and $\Psi : [D \to D] \to D$ are continuous isomorphisms.

The actual construction need not detain us since we are only concerned with the existence of such a model. From now on we shall work with a fixed Scott model $\boldsymbol{D}$.

The semantics is given relative to an assignment function g which assigns elements of D to variables and an interpretation function i which assigns elements of D to constants. We shall employ the notation $g(d/x)$ for that assignment function identical to g except that d is bound to x. We drop all reference to $\boldsymbol{D}$ in the semantic definition which follows.

(i1) $\mathrm{I}[x]_g = g(x)$
(i2) $\mathrm{I}[c]_g = i(c)$
(i3) $\mathrm{I}[\lambda x.t]_g = \Psi(\lambda 0.\mathrm{I}[t]_{g(0/x)})$
(i4) $\mathrm{I}[t(t')]_g = \Phi(\mathrm{I}[t]_g)(\mathrm{I}[t']_g)$.

These clauses are all standard and we shall not pause to explain them. The important point is that the functions $\lambda 0.\mathrm{I}[t]_{g(0/x)}$ are all members of the continuous function space $[D \to D]$ and so the definition is sound. Again, we refer the reader to Hindley and Seldin [7] and Barendregt [2] for details concerning Lambda Calculus models.

2.2 The language of wff The language of wff (L) has three types of atomic wff's: $t = s$, $\mathrm{T}(t)$, and $\mathrm{F}(t)$. The first is equality of terms, the second asserts that a term is true, and the third asserts that a term is false. Terms are those of the Lambda Calculus.

Inductive Definition of wff
(iv) If t is a term then $\mathrm{T}(t)$ and $\mathrm{F}(t)$ are wff's
(v) If t and t' are terms then $t = t'$ is a wff
(vi) If A and B are wff's then so are A & B, A ∨ B, ~A, A → B
(vii) If x is a variable and A is a wff then $\forall xA$ and $\exists xA$ are wff's.

We shall employ $\perp$ as an abbreviation for $0 = 1$ and A ↔ B as an abbreviation for A → B & B → A. We also adopt some standard axiomatization of classical logic. This brings us to the notion of a model for the language.

Definition 2.2.1 A model for L is $\boldsymbol{M} = \langle \boldsymbol{D}, T, F \rangle$ where $\boldsymbol{D}$ is a model of the Lambda Calculus, $T: D \rightarrow \{0,1\}$, and $F: D \rightarrow \{0,1\}$, and where for no d in D do we have $T(d) = 1$ and $F(d) = 1$. (The last clause simply insists that no object can be both true and false.)

The wff's of the language L can now be given truth conditions in the standard way, where T and F respectively provide the extensions of the truth and falsity predicates.

$\boldsymbol{M} \models_g s = t$ iff $\mathrm{I}[t]_g = \mathrm{I}[s]_g$
$\boldsymbol{M} \models_g \mathrm{T}(t)$ iff $T(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \models_g \mathrm{F}(t)$ iff $F(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \models_g A \ \& \ B$ iff $\boldsymbol{M} \models_g A$ and $\boldsymbol{M} \models_g B$
$\boldsymbol{M} \models_g A \vee B$ iff $\boldsymbol{M} \models_g A$ or $\boldsymbol{M} \models_g B$
$\boldsymbol{M} \models_g A \rightarrow B$ iff $\boldsymbol{M} \models_g A$ implies $\boldsymbol{M} \models_g B$
$\boldsymbol{M} \models_g \sim A$ iff not $\boldsymbol{M} \models_g A$
$\boldsymbol{M} \models_g \forall x A$ iff for all d in D, $\boldsymbol{M} \models_{g(d/x)} A$
$\boldsymbol{M} \models_g \exists x A$ iff for some d in D, $\boldsymbol{M} \models_{g(d/x)} A$.

Definition 2.2.2 A wff A of L is *valid in a model* $\boldsymbol{M}$ iff $\boldsymbol{M} \models_g A$ for all assignment functions g.

We write $\mathrm{LC} \vdash A$ if A is provable in first-order logic with equality from the axioms of the Lambda Calculus.

2.3 The paradoxes and the Tarski biconditionals The theory we have at present is perfectly harmless since there are no axioms for the truth predicate. We first need to be able to treat wff's as terms so that the truth predicate T can be applied to them. We therefore add a new clause to the language:

(vii) If A is a wff then $\hat{}A$ is a term.

Actually, we do not have to add this as a new clause since we can achieve the same effect by coding. Using the pairing combinator and numerals of the Lambda Calculus we can code the wff's as terms of the Lambda Calculus as follows:

$\hat{}(x) = x$
$\hat{}(c) = c$
$\hat{}(ts) = \hat{}(t)\hat{}(s)$
$\hat{}(\lambda x.t) = \lambda x.\hat{}(t)$
$\hat{}(t = s) = \langle 0, \hat{}(t), \hat{}(s) \rangle$
$\hat{}(\mathrm{T}(t)) = \langle 1, \hat{}t \rangle$
$\hat{}(\mathrm{F}(t)) = \langle 2, \hat{}t \rangle$
$\hat{}(\sim(A)) = \langle 3, \hat{}A \rangle$
$\hat{}(A \ \& \ B) = \langle 4, \hat{}A, \hat{}B \rangle$
$\hat{}(A \vee B) = \langle 5, \hat{}A, \hat{}B \rangle$
$\hat{}(A \rightarrow B) = \langle 6, \hat{}A, \hat{}B \rangle$
$\hat{}(\forall x A) = \langle 7, x, \hat{}A \rangle$
$\hat{}(\exists x A) = \langle 8, x, \hat{}A \rangle$.

The details of this coding are not important. The only important point is that $\hat{}(A)$ and $\hat{}(B)$ will be the same only when A and B are identical wff's, i.e., the representation enjoys a certain *independence property*. Moreover, this property is inherited by the models, i.e., if $I[\hat{}(A)]_g = I[\hat{}(B)]_g$ then A and B will be the same wff. In the model we let **Code** $= \{I[\hat{}A]_g : A$ is a wff and g an assignment function$\}$.

The intuitive principle which governs the logic of the truth predicate is given by the Tarski biconditionals:

TB $T(A) \leftrightarrow A$, for all wff's A,

where $T(A)$ is an abbreviation for $T(\hat{}A)$. Unfortunately, the theory would then be inconsistent; this stems directly from the fixed-point property of the Lambda Calculus. We shall refer to the following as the *Diagonalization* Lemma.

Theorem 2.3.1 *Let $A(x)$ be any wff whose only free variable is x. Then there is a sentence B such that* $LC \vdash B \leftrightarrow A[\hat{}B/x]$.

Proof: Let $f = \lambda x.\hat{}A(x)$ and $t = Yf$. Then by the fixed-point theorem we have $t = \hat{}A[t/x]$. By the equality rule of the Predicate Calculus we have $A[t/x] \leftrightarrow A[\hat{}A[t/x]/x]$. Finally, put $B = A[t/x]$.

A simple application of this result gives us the paradox of the liar.

Corollary 2.3.2 LC + TB *is inconsistent.*

Proof: Let $A(x) = \sim T(x)$. Then by Theorem 2.3.1 we have $B \leftrightarrow A[\hat{}B/x] \leftrightarrow \sim T(B)$. By the Tarski biconditionals we have a contradiction.

As a matter of interest notice that other paradoxes are derivable in the setting of the Lambda Calculus. The facility for abstraction available in the Lambda Calculus enables the derivation of the *Russell* paradox without explicit appeal to the fixed-point combinator, whereas the *Curry* paradox again uses the fixed-point property.

Russell: Write $\{x : B\}$ for $\lambda x.B$ and $x \in y$ for $T(xy)$. Let $t = \{x : \sim(x \in x)\}$ and then put $A = (t \in t)$. Now tt is equal by β-reduction to $\sim(t \in t)$. By the equality rule we have $T(tt) \leftrightarrow T(\sim(t \in t))$, i.e., $A \leftrightarrow T(\sim A)$, which by the Tarski biconditionals yields the equivalence $A \leftrightarrow \sim A$.

Curry: Let $t = Y[\lambda z.T(z) \rightarrow (T(z) \rightarrow T(Z))]$. Then we have $t = T(t) \rightarrow (T(t) \rightarrow T(Z)) = (T(t) \rightarrow (T(t) \rightarrow T(Z))) \rightarrow (T(t) \rightarrow T(Z))$. Hence by the equality rules we have $T(t) \leftrightarrow T(T(t) \rightarrow (T(t) \rightarrow T(Z))) \leftrightarrow T((T(t) \rightarrow (T(t) \rightarrow T(Z))) \rightarrow (T(t) \rightarrow T(Z)))$. Notice that $(T(t) \rightarrow (T(t) \rightarrow T(Z))) \rightarrow (T(t) \rightarrow T(Z))$ is a tautology and so by the Tarski schema we have $T((T(t) \rightarrow (T(t) \rightarrow T(Z))) \rightarrow (T(t) \rightarrow T(Z)))$ and hence $T(T(t) \rightarrow (T(t) \rightarrow T(Z)))$ and $T(t)$. Moreover, we have $T(t) \rightarrow (T(t) \rightarrow T(Z))$ by the Tarski biconditionals. Consequently, we obtain by modus ponens $T(t) \rightarrow T(Z)$ and finally $T(Z)$.

Given that the theory based on TB is inconsistent the urgent questions are what principles T can satisfy and under what circumstances we can maintain the Tarski biconditionals. Different answers to these questions will lead to different theories of truth.

3 A theory of truth and propositions The first theory we shall consider is due to Scott [10] and Aczel [1]. The central notion is Aczel's concept of a *Frege structure*. These structures are models of the Lambda Calculus together with two distinguished subsets—a set of *propositions* and a subset of this set called *truths*. In addition, such structures come equipped with the usual logical constants together with rules for building propositions from such constants and rules for their associated truth conditions. We shall not actually consider Aczel's set-theoretic models but concentrate rather on an axiomatization of such structures. The theory is stated in terms of two basic predicates: the truth predicate T and a second predicate P, where intuitively $P(t)$ asserts that t is a *proposition*. The truth predicate obtains its correct interpretation only on those objects which are propositions. In the present context we define $P(t) =_{def} T(t) \vee F(t)$, i.e., propositions are those objects which are *true* or *false*. The axioms of the theory SA (the axioms of a Frege structure) are given in two parts, corresponding to those for truth and those for propositions.

Axioms of Propositions

(i) $P(A) \ \& \ P(B) \rightarrow P(A \ \& \ B)$
(ii) $P(A) \ \& \ P(B) \rightarrow P(A \vee B)$
(iii) $P(A) \ \& \ (T(A) \rightarrow P(B)) \rightarrow P(A \rightarrow B)$
(iv) $P(A) \rightarrow P(\sim A)$
(v) $\forall x P(A) \rightarrow P(\forall x A)$
(vi) $\forall x P(A) \rightarrow P(\exists x A)$
(vii) $P(s = t)$.

Axioms of Truth

(i) $P(A) \ \& \ P(B) \rightarrow (T(A \ \& \ B) \leftrightarrow T(A) \ \& \ T(B))$
(ii) $P(A) \ \& \ P(B) \rightarrow (T(A \vee B) \leftrightarrow T(A) \vee T(B))$
(iii) $P(A) \ \& \ (T(A) \rightarrow P(B)) \rightarrow (T(A \rightarrow B) \leftrightarrow (T(A) \rightarrow T(B)))$
(iv) $P(A) \rightarrow (T(\sim A) \leftrightarrow \sim T(A))$
(v) $\forall x P(A) \rightarrow (T(\forall x A) \leftrightarrow \forall x T(A))$
(vi) $\forall x P(A) \rightarrow (T(\exists x A) \leftrightarrow \exists x T(A))$
(vii) $T(s = t) \leftrightarrow s = t$
(viii) $\sim (T(A) \ \& \ F(A))$.

The theory does not assign the standard Tarski truth conditions to all the wff's but only those which are provably propositions. Moreover, the structure of propositions is *predicative* in that we can establish that something is a proposition only by proving that its subformulas denote propositions. The axioms for truth are then the standard Tarski ones, where propositions are the objects of the truth predicate. In an important sense these axioms reflect the minimal conditions one would expect of any theory of truth. The other theories we shall consider all have the above theory as a consequence.

3.1 Models of **SA** Our set-theoretic models of SA are given by a slight variation on Aczel's construction, one that is similar to the account of Scott [10]. This is done largely to compare this theory with that of Kripke–Feferman–Gilmore we shall consider in the next section. To construct the models we first

give a different semantics for L, namely that of Kleene (strong) three-valued logic. We define two semantic relations $\vdash$ (true) and $\dashv$ (false) by simultaneous recursion as follows:

The strong Kleene truth conditions for L

$\boldsymbol{M} \vdash_g s = t$ iff $\mathrm{I}[t]_g = \mathrm{I}[s]_g$
$\boldsymbol{M} \vdash_g \mathrm{T}(t)$ iff $T(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \vdash_g \mathrm{F}(t)$ iff $F(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \vdash_g A \,\&\, B$ iff $\boldsymbol{M} \vdash_g A$ and $\boldsymbol{M} \vdash_g B$
$\boldsymbol{M} \vdash_g A \vee B$ iff $\boldsymbol{M} \vdash_g A$ or $\boldsymbol{M} \vdash_g B$
$\boldsymbol{M} \vdash_g A \rightarrow B$ iff $\boldsymbol{M} \dashv_g A$ or $\boldsymbol{M} \vdash_g B$
$\boldsymbol{M} \vdash_g \sim A$ iff $\boldsymbol{M} \dashv_g A$
$\boldsymbol{M} \vdash_g \forall x A$ iff for all d in D, $\boldsymbol{M} \vdash_{g(d/x)} A$
$\boldsymbol{M} \vdash_g \exists x A$ iff for some d in D, $\boldsymbol{M} \vdash_{g(d/x)} A$.

$\boldsymbol{M} \dashv_g s = t$ iff $\mathrm{I}[t]_g \neq \mathrm{I}[s]_g$
$\boldsymbol{M} \dashv_g \mathrm{T}(t)$ iff $F(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \dashv_g \mathrm{F}(t)$ iff $T(\mathrm{I}[t]_g) = 1$
$\boldsymbol{M} \dashv_g A \,\&\, B$ iff $\boldsymbol{M} \dashv_g A$ or $\boldsymbol{M} \dashv_g B$
$\boldsymbol{M} \dashv_g A \vee B$ iff $\boldsymbol{M} \dashv_g A$ and $\boldsymbol{M} \dashv_g B$
$\boldsymbol{M} \dashv_g A \rightarrow B$ iff $\boldsymbol{M} \vdash_g A$ and $\boldsymbol{M} \dashv_g B$
$\boldsymbol{M} \dashv_g \sim A$ iff $\boldsymbol{M} \vdash_g A$
$\boldsymbol{M} \dashv_g \forall x A$ iff for some d in D, $\boldsymbol{M} \dashv_{g(d/x)} A$
$\boldsymbol{M} \dashv_g \exists x A$ iff for all d in D, $\boldsymbol{M} \dashv_{g(d/x)} A$.

The main theorem for the construction of models for SA employs the *monotonicity* property of the above truth conditions.

Definition 3.1.1 Let $\boldsymbol{M} = \langle \boldsymbol{D}, T, F \rangle$ and $\boldsymbol{N} = \langle \boldsymbol{D}, T', F' \rangle$ be models. Define $\boldsymbol{M} \subseteq \boldsymbol{N}$ iff $(\forall d \in D)(T(d) = 1 \rightarrow T'(d) = 1 \;\&\; F(d) = 1 \rightarrow F'(d) = 1)$. Define $\boldsymbol{M} \equiv \boldsymbol{N}$ iff $\boldsymbol{M} \subseteq \boldsymbol{N}$ and $\boldsymbol{N} \subseteq \boldsymbol{M}$.

In order to construct a model for SA we revise the extensions of truth and falsity in an attempt to force the Tarski biconditionals. Let $\boldsymbol{M} = \langle \boldsymbol{D}, T, F \rangle$ be a model for L. Define $\boldsymbol{M}' = \langle \boldsymbol{D}, T', F' \rangle$, where

$$T'(\mathrm{I}[A]_g) = 1 \text{ iff } \boldsymbol{M} \vdash_g A$$
$$F'(\mathrm{I}[A]_g) = 1 \text{ iff } \boldsymbol{M} \dashv_g A.$$

On those elements of D which are not elements of **Code**, T and F do not change. This definition is legitimate since the $\hat{}$ function enjoys the previously mentioned independence property.

Theorem 3.1.2 *Let $\boldsymbol{M}$ and $\boldsymbol{N}$ be two models of L. Then $\boldsymbol{M} \subseteq \boldsymbol{N}$ implies $\boldsymbol{M}' \subseteq \boldsymbol{N}'$.*

Proof: By induction on wff's we show that $\boldsymbol{M} \vdash_g A$ implies $\boldsymbol{N} \vdash_g A$ and $\boldsymbol{M} \dashv_g A$ implies $\boldsymbol{N} \dashv_g A$, where the atomic cases follow directly from the assumption.

Using this basic step of revision we can define an ordinal sequence of truth and falsity predicates: $T(\alpha)$, $F(\alpha)$ and models $\boldsymbol{M}(\alpha)$ for $\alpha \geq 0$ where

$T(0) = \perp\!\!\!\perp$, *where* $(\forall d \in D)\ (\perp\!\!\!\perp(d) = 0)$
$F(0) = \perp\!\!\!\perp$
$T(\alpha + 1) = T(\alpha)'$
$F(\alpha + 1) = F(\alpha)'$
$T(\delta)(d) = 1$ iff $(\exists a < \delta)\forall\beta(\alpha \leq \beta < \delta)(T(\beta)(d) = 1)$, for limit ordinal δ
$F(\delta)(d) = 1$ iff $(\exists a < \delta)\forall\beta(\alpha \leq \beta < \delta)(F(\beta)(d) = 1)$, for limit ordinal δ.

The following *fixed-point* theorem facilitates the construction of models for SA.

Theorem 3.1.3 *There is a model* $\boldsymbol{M}^* = \langle \boldsymbol{D}, T^*, F^* \rangle$ *such that* $\boldsymbol{M}^* \equiv (\boldsymbol{M}^*)'$.

Proof: Using the monotonicity of the $'$ operation we prove by induction that for all ordinals β and μ if $\beta < \mu$ then $\boldsymbol{M}(\beta) \subseteq \boldsymbol{M}(\mu)$. It then follows that there exists a least δ such that $\boldsymbol{M}(\delta) = \boldsymbol{M}(\delta)'$. Let $\boldsymbol{M}^*$ be this model. Then it is easy to see that $T^*(\mathrm{I}[A]_g) = 1$ iff $\boldsymbol{M}^* \vdash_g A$, and $F^*(\mathrm{I}[A]_g) = 1$ iff $\boldsymbol{M}^* \dashv_g A$.

Theorem 3.1.4 $\boldsymbol{M}^*$ *is a model of* SA.

Proof: The axioms of SA are then automatic from the Kleene truth conditions.

The elegant theory SA forms a core for all the theories we shall consider. In fact the model $\boldsymbol{M}^*$ supports a stronger theory of truth, one which satisfies the Kleene truth conditions exactly.

4 A theory of truth and falsity The second theory we shall consider is essentially classical and is a version of the theory of Kripke [9]. The theory is due to Gilmore [4] and Feferman [3]. The theory, although classical in its *external* logic (i.e., the logic of wff's is classical), has a residue of three-valued logic in its *internal* logic of the truth predicate.

4.1 The axioms of the theory **KFG** The theory is stated in terms of certain principles which govern the logic of the predicates T and F.

Axioms of KFG

A1 $\mathrm{T}(t) \leftrightarrow \mathrm{T}(\mathrm{T}(t))$
A2 $\mathrm{F}(t) \leftrightarrow \mathrm{F}(\mathrm{T}(t))$
A3 $\mathrm{F}(t) \leftrightarrow \mathrm{T}(\mathrm{F}(t))$
A4 $\mathrm{T}(t) \leftrightarrow \mathrm{F}(\mathrm{F}(t))$
A5 $\mathrm{T}(t = s) \leftrightarrow t = s$
A6 $\mathrm{T}(t \neq s) \leftrightarrow t \neq s$
DIS $\mathrm{T}(A) \rightarrow \mathrm{C}(A)$, where $\mathrm{C}(A) =_{def} {\sim}\mathrm{F}(A)$
F1 $\mathrm{T}(A\ \&\ B) \leftrightarrow \mathrm{T}(A)\ \&\ \mathrm{T}(B)$
F2 $\mathrm{F}(A\ \&\ B) \leftrightarrow \mathrm{F}(A) \vee \mathrm{F}(B)$
F3 $\mathrm{T}(A \vee B) \leftrightarrow \mathrm{T}(A) \vee \mathrm{T}(B)$
F4 $\mathrm{F}(A \vee B) \leftrightarrow \mathrm{F}(A)\ \&\ \mathrm{F}(B)$
F5 $\mathrm{T}({\sim}A) \leftrightarrow \mathrm{F}(A)$
F6 $\mathrm{F}({\sim}A) \leftrightarrow \mathrm{T}(A)$
F7 $\mathrm{T}(A \rightarrow B) \leftrightarrow \mathrm{T}(B) \vee \mathrm{F}(A)$

F8 $F(A \to B) \leftrightarrow T(A)\ \&\ F(B)$
F9 $T(\forall xA) \leftrightarrow \forall xT(A)$
F10 $F(\forall xA) \leftrightarrow \exists xF(A)$
F11 $T(\exists xA) \leftrightarrow \exists xT(A)$
F12 $F(\exists xA) \leftrightarrow \forall xF(A)$.

The first group of axioms (A1–A6) concern the behavior of atomic sentences under T and F and allow the derivation of the Tarski biconditionals for atomic assertions. The main group of axioms for truth and falsity (F1–F12) are exactly the expression of the truth conditions of Kleene (strong) three-valued logic. We alluded to this earlier: the *internal* logic of truth is three-valued.

Theorem 4.1.1 *The model* $\boldsymbol{M}^* = \langle \boldsymbol{D}, T^*, F^* \rangle$ *is a model of the theory* KFG.

Proof: The model constructed for SA serves as a model for KFG. The soundness of the axioms is automatic from the Kleene truth conditions.

The following is almost immediate:

Theorem 4.1.2 *The theory* SA *is derivable in* KFG.

In addition we have the following equivalences:

Theorem 4.1.3 *The following are provable in* KFG*:*
(i) $P(A) \leftrightarrow P(T(A))$
(ii) $P(A) \leftrightarrow T(P(A))$
(iii) $P(A) \leftrightarrow (T(A) \leftrightarrow A\ \&\ T(\sim A) \leftrightarrow \sim A)$.

Proof: All are straightforward.

We can also prove the following derived rules of inference which govern the truth predicate.

A Natural Deduction System for Truth

Conjunction

$$\frac{T(A)\quad T(B)}{T(A\ \&\ B)} \qquad \frac{T(A\ \&\ B)}{T(A)} \qquad \frac{T(A\ \&\ B)}{T(B)}$$

Implication

$$\frac{\begin{array}{c}[T(A)]\\ \vdots \\ T(B)\end{array}\quad P(A)}{T(A \to B)} \qquad \frac{T(A)\quad T(A \to B)}{T(B)}$$

Disjunction

$$\frac{T(A)}{T(A \vee B)} \qquad \frac{T(B)}{T(A \vee B)} \qquad \frac{T(A \vee B)\quad \begin{array}{c}[T(A)]\\ \vdots \\ T(C)\end{array}\quad \begin{array}{c}[T(B)]\\ \vdots \\ T(C)\end{array}}{T(C)}$$

Negation

$$\dfrac{\begin{array}{c}[\mathrm{T}(A)]\\ \vdots\\ \bot \quad \mathrm{P}(A)\end{array}}{\mathrm{T}(\sim A)} \qquad \dfrac{\mathrm{T}(A) \quad \mathrm{T}(\sim A)}{\bot}$$

Classical Negation

$$\dfrac{\begin{array}{c}[\mathrm{T}(\sim A)]\\ \vdots\\ \bot \quad \mathrm{P}(A)\end{array}}{\mathrm{T}(A)}$$

Universal Quantification

$$\dfrac{\mathrm{T}(A[c])}{\mathrm{T}(\forall x A)} \qquad \dfrac{\mathrm{T}(\forall x A)}{\mathrm{T}(A[t])}$$

Existential Quantification

$$\dfrac{\mathrm{T}(A[t])}{\mathrm{T}(\exists x A)} \qquad \dfrac{\mathrm{T}(\exists x A) \quad \begin{array}{c}[\mathrm{T}(A[c])]\\ \vdots\\ \mathrm{T}(C)\end{array}}{\mathrm{T}(C)}$$

We assume the normal side-conditions on the quantifier rules for existential elimination and universal introduction. Observe that there are side-conditions on propositions in the implication introduction rule whereas there are no such conditions for disjunction or existential elimination.

Theorem 4.1.4 *The above rules are derivable in* KFG.

Proof: Once again the details are tedious but simple to check.

4.2 A reformulation The original formulation of the theory was not given in terms of the above axioms A1–A6, DIS, and F1–F12 but in terms of these schemas:

C1 $\mathrm{T}(A) \leftrightarrow A^+$
C2 $\mathrm{F}(A) \leftrightarrow A^-$

where A^+ and A^- are defined as follows:

Definition 4.2.1 Let A be any wff of L. Then A^+ and A^- are defined by recursion:

(i) If A is $t = s$ then $A^+ = A$ and $A^- = t \neq s$
(ii) If A is $\mathrm{T}(t)$ then $A^+ = A$ and $A^- = \mathrm{F}(t)$
If A is $\mathrm{F}(t)$ then $A^+ = A$ and $A^- = \mathrm{T}(t)$
(iii) If A is $\sim B$ then $A^+ = B^-$ and $A^- = B^+$
(iv) If A is $B \,\&\, C$ then $A^+ = B^+ \,\&\, C^+$ and $A^- = B^- \vee C^-$
(v) If A is $B \vee C$ then $A^+ = B^+ \vee C^+$ and $A^- = B^- \,\&\, C^-$
(vi) If A is $\forall x B$ then $A^+ = \forall x B^+$ and $A^- = \exists x B^-$
(vii) If A is $\exists x B$ then $A^+ = \exists x B^+$ and $A^- = \forall x B^-$
(viii) If A is $B \rightarrow C$ then $A^+ = B^- \vee C^+$ and $A^- = B^+ \,\&\, C^-$

In essence these approximations are the result of pushing all negations into atomic position and replacing all such negations by *internal negations.*

Lemma 4.2.2 *For each wff A we have:*
(i) $A^+ \to A$
(ii) $A^- \to \sim A$
(iii) $(A \leftrightarrow A^+) \to (\mathrm{T}(A) \leftrightarrow A)$
(iv) $(\sim A \leftrightarrow A^-) \to (\mathrm{F}(A) \leftrightarrow \sim A)$
(v) $(A^+ \leftrightarrow A \;\&\; A^- \leftrightarrow \sim A) \leftrightarrow \mathrm{P}(A)$.

Proof: The proofs of (i) and (ii) are routine and are established by simultaneous induction. The third, fourth, and fifth parts are immediate.

Theorem 4.2.3 *The theory* (A1–A6) + DIS + (F1–F12) *is equivalent to* DIS + (C1–C2).

Proof: It is obvious that each of the axioms A1–A4 and F1–F18 follow from C1 and C2. The converse direction is by induction on wff's. Use A1–A6 for the atomic cases and F1–F12 for the induction clauses.

This version of the theory slightly disguises its three-valued origins, and the theory is best summarized by saying that $\mathrm{T}(A)$ means that A is true in all Kleene models.

4.3 **KFG** *as a modal theory* To prepare the route to the next theory we now investigate some more of the consequences of KFG and in particular the derivability of certain modal principles.

Theorem 4.3.1 *The following are provable in* KFG*:*

T $\mathrm{T}(A) \to A$
S4 $\mathrm{T}(A) \to \mathrm{T}(\mathrm{T}(A))$
IP $\mathrm{T}(A \to B) \to (\mathrm{T}(A) \to \mathrm{T}(B))$
BAR $\forall x \mathrm{T}(A) \to \mathrm{T}(\forall x A)$.

Proof: We illustrate the proof for IP. It is sufficient to show that $(A^- \vee B^+) \leftrightarrow (\mathrm{T}(A)^- \vee \mathrm{T}(B)^+)$, which is clear.

T, S4, IP, and BAR are the characteristic axioms of S4 modal logic. As regards the S5 axiom the following is provable:

$$(\mathrm{C}(A) \to \mathrm{T}(\mathrm{C}(A))) \to \mathrm{P}(A).$$

As a consequence S5 is not derivable: its truth renders everything a proposition and hence the theory is inconsistent. However, there are further modal axioms which are provable.

Theorem 4.3.2 *The following are provable in* KFG*:*

S $\mathrm{T}(\mathrm{T}(A) \to A) \leftrightarrow \mathrm{P}(A)$
IPT $\mathrm{T}(\mathrm{T}(A) \to \mathrm{T}(B)) \leftrightarrow \mathrm{T}(A \to B)$
R $\mathrm{T}(A) \to \mathrm{T}(\mathrm{C}(A))$
L $\mathrm{T}(\mathrm{C}(A)) \to \mathrm{T}(A)$.

Proof: Once again these are all easy to verify.

With these axioms in place it appears that KFG supports a distinctive modal theory of truth. But one important factor is missing, namely a rule of necessitation:

$$\text{KFG} \vdash A \text{ implies KFG} \vdash \text{T}(A).$$

This rule is not provable in KFG. Indeed, its addition renders the theory inconsistent. To see this observe that KFG $\vdash A \vee \sim A$, hence by the rule of necessitation KFG $\vdash \text{T}(A \vee \sim A)$ and so by the axiom for disjunction we have $\text{T}(A) \vee \text{F}(A)$, i.e., everything will be a proposition. Indeed this follows from a weak rule of necessitation (one which allows the derivation of $\text{T}(A)$ only when A is provable in classical logic). Although the logic has the appearance of a modal theory, it is not very interesting given the lack of any obvious rule of necessitation. Indeed, the schemas C1 and C2 under the above rule would lead to:

$$\text{T}(\text{T}(A) \leftrightarrow A^{+})$$
$$\text{T}(\text{F}(A) \leftrightarrow A^{-}),$$

but the truth of either again yields that everything is a proposition. This is rather unfortunate since one would like the assertion of the truth of the schemas of the theory to be true. We now turn to a theory which looks more promising as a modal theory.

5 A theory of truth The theory KFG is characterized by a classical external logic and a three-valued internal one. In this section we investigate a theory of truth where the internal logic of the truth predicate is classical. The point of departure is the semantic theory of Gupta–Herzberger.

5.1 The Gupta–Herzberger semantic theory The Gupta–Herzberger approach to the theory of truth is based completely on classical semantics. Wff's are given a classical interpretation both internally and externally. To begin with, we offer an account of the Gupta–Herzberger process of *revision*. We shall be brief since our main concern is with the logics of truth that result. We shall follow the exposition given by Herzberger.

The idea behind this approach to truth is simple enough and resides in the desire to maintain as many instances of the Tarski biconditionals as possible. We shall construct an ordinal sequence of models for L where the extension of the truth predicate is continually revised.

Initially we let $T: D \to \{0,1\}$ and $F: D \to \{0,1\}$ be arbitrary. We revise the extension of truth and falsity in an attempt to force the Tarski schema:

$$T'(\text{I}[A]_g) = 1 \text{ iff } \boldsymbol{M} \models_g A$$
$$F'(\text{I}[A]_g) = 1 \text{ iff } \boldsymbol{M} \models_g \sim A.$$

On those elements of D which are not elements of **Code**, T and F do not change. The important point to observe is that this process of revision is not monotone since $T(d) = 1$ does not imply $T'(d) = 1$.

Using this basic step of revision we can define a sequence of truth and falsity predicates $T(\alpha)$, $F(\alpha)$ for $\alpha \geq 0$ as follows:

(i) $T(0) = T$
$F(0) = F$
(ii) $T(\alpha + 1) = T(\alpha)'$
$F(\alpha + 1) = F(\alpha)'$.

Because the operation of revision is not monotone, in order to carry the process through to transfinite ordinals we cannot simply select the union of all the T's at limit ordinals. Here we follow the lead of Herzberger:

(iii) For limit ordinal δ define:
$T(\delta)(d) = 1$ iff $(\exists\alpha < \delta)\forall\beta(\alpha \leq \beta < \delta)(T(\beta)(d) = 1)$
$F(\delta)(d) = 1$ iff $(\exists\alpha < \delta)\forall\beta(\alpha \leq \beta < \delta)(F(\beta)(d) = 1)$.

It should be pointed out that there are many options concerning the definition of the truth predicate at limit ordinals and no doubt different choices would lead to different theories of truth. Gupta [5] and Herzberger [6] contain indications of the different choices available, but we shall not pursue this here.

Definition 5.1.1 An element d in D is *positively stable* iff $\exists\alpha\forall\beta \geq \alpha$ $(T(\beta)(d) = 1)$; it is *negatively stable* iff $\exists\alpha\forall\beta \geq \alpha(F(\beta)(d) = 1)$. An element d of D is *stable* iff d is positively or negatively stable. We say that d is *positively stable from an ordinal* α (*negatively stable from* α) iff $\forall\beta \geq \alpha(T(\beta)(d) = 1)$ $(\forall\beta \geq \alpha(F(\beta)(d) = 1))$.

Definition 5.1.2 An ordinal σ is a *stabilization ordinal* iff
(i) For each d in D, d is positively stable iff $T(\sigma)(d) = 1$. For each d in D, d is negatively stable iff $F(\sigma)(d) = 1$.
(ii) For each d in D, d is positively (negatively) stable implies that d is positively (negatively) stable from σ.

Stabilization ordinals characterize the stable objects exactly. The central result for our purposes is the following.

Theorem 5.1.3 (Herzberger) *There exists a stabilization ordinal.*

We are primarily interested in those wff's which are valid at such models.

Definition 5.1.4 A wff is *sound* iff it is valid at every stabilization ordinal.

This is only a brief account of the Gupta–Herzberger theory of truth. It is sufficient for our purposes; the two papers referenced contain more details. We shall be concerned with exploring the *logic* of stable truth. In this regard observe that the Gupta–Herzberger semantics for the truth predicate has a modal flavor to it: for a wff A to be stably true A must be true in all models after some point in the revision process. Moreover, at stabilization ordinals $\mathrm{T}(A)$ means that A is stably true and $\mathrm{F}(A)$ that A is stably false. In the rest of the paper we explore this modal interpretation of the truth predicate and consider various modal logics all consistent with this stability interpretation.

5.2 A simple deontic logic of truth The weakest system we shall consider is not really capable of being interpreted as a logic of necessity and possibility. It is a weak deontic logic, whose interest lies in its connection with Frege structures.

The Theory M

DIS $T(A) \rightarrow C(A)$, where $C(A) =_{def} \sim F(A)$
IP $T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B))$
BAR $\forall x T(A) \rightarrow T(\forall x A)$
NEC If $LC \vdash A$ then $T(A)$ is a thesis of M.

This is a very weak modal theory. In particular, it has a weakened rule of necessitation. One can only conclude that $T(A)$ if A is derivable from the underlying theory of the Lambda Calculus LC. The axiom IP is self-explanatory. The axiom BAR is essentially the Barcan formula. The axiom DIS, familiar from deontic logic, prevents both $T(A)$ and $T(\sim A)$ from being simultaneously true.

We first establish the soundness of this logic under the stability interpretation, where by *soundness* we mean that all its theorems are sound.

Theorem 5.2.1 *If* $M \vdash A$ *then* A *is sound.*

Proof: The axiom IP follows because modus ponens preserves stability: if $T(A)$ and $T(A \rightarrow B)$ are true at a stabilization ordinal then A and $A \rightarrow B$ will be stably true and so will B. Hence $T(B)$ will be true at this stabilization ordinal. Next consider the Barcan formula. This follows because the domain of individuals is fixed throughout the revision process. The inference rule of weak necessitation NEC preserves soundness, because any wff provable in the underlying theory of the Lambda Calculus will be true in all models and will thus be stably true. DIS is true at any stabilization ordinal because $T(A)$ means that A will be true from the stabilization point onwards, and so there is no possibility of A being false let alone stably false.

Theorem 5.2.2 *The following are derivable in* M:

(i) $T(A \ \& \ B) \leftrightarrow T(A) \ \& \ T(B)$
(ii) $F(A) \vee F(B) \rightarrow F(A \ \& \ B)$
(iii) $T(A) \vee T(B) \rightarrow T(A \vee B)$
(iv) $F(A \vee B) \leftrightarrow F(A) \ \& \ F(B)$
(v) $T(\sim A) \leftrightarrow F(A)$
(vi) $F(\sim A) \leftrightarrow T(A)$
(vii) $T(B) \vee F(A) \rightarrow T(A \rightarrow B)$
(viii) $F(A \rightarrow B) \leftrightarrow T(A) \ \& \ F(B)$
(ix) $T(\forall x A) \leftrightarrow \forall x T(A)$
(x) $\exists x F(A) \rightarrow F(\forall x A)$
(xi) $\exists x T(A) \rightarrow T(\exists x A)$
(xii) $F(\exists x A) \leftrightarrow \forall x F(A)$.

Proof: These are all quite straightforward and follow from the appropriate classical truths and judicious applications of IP, BAR, and NEC. We illustrate with (i). From left to right we employ the tautology $A \ \& \ B \rightarrow A$, NEC, and IP; from right to left we employ the tautology $A \rightarrow (B \rightarrow A \ \& \ B)$, NEC, and two applications of IP.

The converse directions of (ii), (iii), (vii), (x), and (xi) are not generally derivable. In fact, the theory which results from the inclusion of these converse directions gives precisely the strong Kleene truth conditions for T and F, which is essentially the theory KFG of Gilmore [4]–Feferman [3].

By way of further unpacking the content of the theory M we consider the theory SA of Aczel's Frege structures. In the present context we again define $P(t) =_{def} T(t) \vee F(t)$, i.e., propositions are those objects which are true or false. Under the present interpretation propositions are those objects which are stable.

Theorem 5.2.3 *The axioms of a Frege structure are derivable in* M.

Proof: For propositions we illustrate with (v). The assumption yields $\forall x(T(Ax) \vee T(\sim Ax))$. This implies $\forall x T(Ax) \vee \exists x T(\sim Ax)$. By BAR we obtain $T(\forall x A) \vee \exists x T(\sim Ax)$. By the classical truth $\sim At \rightarrow \exists x \sim A$, NEC, and IP we obtain $T(\forall x A) \vee T(\sim \forall x A)$, as required. For the axioms of T consider the case of negation. $T(\sim A) \rightarrow \sim T(A)$ follows from the axiom DIS whereas $\sim T(A) \rightarrow T(\sim A)$ follows only because $P(A)$.

We can also derive a natural deduction formulation of a logic of truth.

A Natural Deduction System For Truth

Conjunction

$$\frac{T(A) \quad T(B)}{T(A \,\&\, B)} \qquad \frac{T(A \,\&\, B)}{T(A)} \qquad \frac{T(A \,\&\, B)}{T(B)}$$

Implication

$$\frac{\begin{array}{c}[T(A)] \\ \vdots \\ T(B) \quad P(A)\end{array}}{T(A \rightarrow B)} \qquad \frac{T(A) \quad T(A \rightarrow B)}{T(B)}$$

Disjunction

$$\frac{T(A)}{T(A \vee B)} \qquad \frac{P(A) \quad T(A \vee B) \quad \begin{array}{c}[T(A)] \\ \vdots \\ T(C)\end{array} \quad \begin{array}{c}[T(B)] \\ \vdots \\ T(C)\end{array}}{T(C)}$$

$$\frac{T(B)}{T(A \vee B)} \qquad \frac{P(B) \quad T(A \vee B) \quad \begin{array}{c}[T(A)] \\ \vdots \\ T(C)\end{array} \quad \begin{array}{c}[T(B)] \\ \vdots \\ T(C)\end{array}}{T(C)}$$

Negation

$$\frac{\begin{array}{c}[T(A)] \\ \vdots \\ \bot \quad P(A)\end{array}}{T(\sim A)} \qquad \frac{T(A) \quad T(\sim A)}{\bot}$$

Classical Negation

$$\frac{\begin{array}{c}[T(\sim A)] \\ \vdots \\ \bot \quad P(A)\end{array}}{T(A)}$$

$$\text{Universal Quantification}\quad \frac{\mathrm{T}(A[c])}{\mathrm{T}(\forall xA)} \qquad \frac{\mathrm{T}(\forall xA)}{\mathrm{T}(A[t])}$$

$$\text{Existential Quantification}\quad \frac{\mathrm{T}(A[t])}{\mathrm{T}(\exists xA)} \qquad \frac{\forall x\mathrm{P}(A) \quad \mathrm{T}(\exists xA) \quad \begin{array}{c}[\mathrm{T}(A[c])]\\ \vdots \\ \mathrm{T}(C)\end{array}}{\mathrm{T}(C)}$$

We assume the normal side-conditions on the quantifier rules for existential elimination and universal introduction. These rules provide us with a natural deduction version of a logic of stable truth. The rules are not carbon copies of the rules for the classical predicate calculus because of the additional conditions regarding stability/propositions in the negation rules, the implication introduction rule, and the disjunction and existential elimination rules. Observe the differences with these rules and those of the system KFG; in the latter there are no side-conditions on propositions for the disjunction and existential elimination rules. In this respect the above system is more uniform.

Theorem 5.2.4 *The above rules are derivable in* M.

Proof: The details are tedious but simple to check. We illustrate with the rule for implication introduction. Given $\mathrm{P}(A)$ there are two possibilities, $\mathrm{T}(A)$ or $\mathrm{T}(\sim A)$. Assume that $\mathrm{T}(A)$. Use the tautology $B \rightarrow (A \rightarrow B)$ and NEC to derive $\mathrm{T}(B \rightarrow (A \rightarrow B))$. Under the assumption that $\mathrm{T}(A)$ we can derive from the assumptions of the rule that $\mathrm{T}(B)$. Now employ IP to derive $\mathrm{T}(A \rightarrow B)$. If $\mathrm{T}(\sim A)$ then use the tautology $\sim A \rightarrow (A \rightarrow B)$, NEC, and IP to get $\mathrm{T}(A \rightarrow B)$.

This is no doubt a quite interesting logic of truth but from a modal perspective the above theory is a very weak one; so what happens when we try to strengthen it? There are two obvious ways of achieving this: one corresponds to the addition of further modal axioms and the other to the rule of necessitation. We consider the latter first.

5.3 The modal logic **D** The modal logic which results from permitting the full axiom of necessitation is the quantifier version of the classical deontic logic D (plus Barcan). This is defined by the following axioms and rules:

The Logic D
DIS $\mathrm{T}(A) \rightarrow \mathrm{C}(A)$
IP $\mathrm{T}(A \rightarrow B) \rightarrow (\mathrm{T}(A) \rightarrow \mathrm{T}(B))$
BAR $\forall x\mathrm{T}(A) \rightarrow \mathrm{T}(\forall xA)$
NEC′ If $\mathrm{D} \vdash A$ then $\mathrm{D} \vdash \mathrm{T}(A)$.

To establish the soundness of D under the stability interpretation we first establish that each of the axioms DIS, IP, and BAR is not just sound (e.g., DIS is true at stabilization ordinals) but stably true (e.g., T(DIS) is true at stabilization ordinals). We thus need to establish the soundness of the following axioms:

SDIS $T(T(A) \rightarrow C(A))$
SIP $T(T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B)))$
SBAR $T(\forall x T(A) \rightarrow T(\forall x A))$.

Theorem 5.3.1 SIP, SDIS, *and* SBAR *are sound.*

Proof: For SIP we have to show that $T(T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B)))$ is true at any stabilization ordinal. First note that $T(A \rightarrow B) \rightarrow (T(A) \rightarrow T(B))$ is true at any successor ordinal, by the definition of revision. Moreover, at limit ordinals if $A \rightarrow B$ has been true from some ordinal less than the limit ordinal, and A likewise, then B must have been true from the greater of the two ordinals, and thus is true at the limit. For the stability of DIS we only have to observe that $T(A)$ always excludes the possibility of $T(\sim A)$ at both successor ordinals and limits. The argument for SBAR again relies on the constancy of the domain of individuals.

Theorem 5.3.2 *If* $D \vdash A$ *then* A *is stably true.*

Proof: We establish the result by induction on the proofs in D. First observe that all the proof rules of the classical Predicate Calculus preserve stability. If A is an instance of any of the axioms of D then the result follows from the previous theorem. Finally consider NEC′. Here we only have to observe that if a wff is stably true then it is stably true that it is stably true. This follows essentially from the definition of revision at successor ordinals.

Corollary 5.3.3 *If* $M \vdash A$ *then* A *is stably true.*

In the theory D we have a full principle of substitution:

Sub $A \leftrightarrow B \rightarrow \Psi[A] \leftrightarrow \Psi[B]$

where Ψ is any context in which a wff can be meaningfully substituted. This is easy to prove by induction on the context. The full rule of necessitation plays the crucial role where the context is T itself.

The main result of this section is that the modal logic D is a consistent logic of truth, and moreover all the theorems of D are stably true. The logic D is thus a logic of stable truth.

5.4 Further modal axioms The axioms of the standard systems of modal logic (**T, S4, S5**) are the next target for the strengthening of our logic of truth. Here there are some real surprises in comparison with standard modal systems. We consider various additional modal axioms beginning with those for **T**, **S4**, and **S5.**

In the present context the characteristic axiom for the modal logic **T** is:

T $T(A) \rightarrow A$.

Theorem 5.4.1 T *is sound.*

Proof: At stabilization ordinals $T(A)$ states that A is stably true. Observe that if A is stably true then A is true at a stabilization ordinal. This follows because of the rule of revision at successor ordinals and the fact that if $T(A)$ is true at a stabilization ordinal then $T(A)$ will be true at its successor.

Of course, the axiom T implies DIS, but whereas DIS is stably true, T is not. Indeed, if it were then the modal logic **T** would be a consistent logic of truth, but we have:

Theorem 5.4.2 *The modal logic* **T** *is inconsistent as a logic of truth.*

Proof: Let $t = Y(\lambda x.{\sim}\mathrm{T}(x))$ and $A = \mathrm{T}(t)$. Then by the fixed-point property of Y and the equality rule of the Predicate Calculus we have $A \leftrightarrow \mathrm{T}({\sim}A)$. Assume that $\mathrm{T}({\sim}A)$, then by the **T**-axiom we obtain ${\sim}A$. By the equivalence $A \leftrightarrow \mathrm{T}({\sim}A)$ we can conclude that ${\sim}\mathrm{T}({\sim}A)$. But now we have a contradiction to the assumption $\mathrm{T}({\sim}A)$ and hence ${\sim}\mathrm{T}({\sim}A)$. So by the equivalence $A \leftrightarrow \mathrm{T}({\sim}A)$, ${\sim}A$ is a theorem of the modal logic **T**. This is not yet a contradiction, but ${\sim}A$ leads by strong necessitation to $\mathrm{T}({\sim}A)$, which we know leads to a contradiction.

This result is essentially the truth-theoretic version of the result of Montague & Kaplan [8]. Their derivation is based upon the *Hangman* paradox. In conclusion, we can add T as an axiom, but it must not enter into proofs involving the rule of necessitation.

In regard to the stable truth of the **T**-axiom the following principle is sound:

S $\mathrm{T}(\mathrm{T}(A) \rightarrow A) \rightarrow \mathrm{P}(A)$.

Theorem 5.4.3 S *is stably true.*

Proof: S will be true at any successor ordinal since its conclusion always will be. So suppose that $\mathrm{T}(\mathrm{T}(A) \rightarrow A)$ is true at some limit ordinal. Then $\mathrm{T}(A) \rightarrow A$ will be true from some ordinal β less than this limit. If A is true at some ordinal greater than β but less than the limit then A will be true at every ordinal greater than β but less than the limit. It follows that $\mathrm{T}(A)$ will be true at the limit; otherwise $\mathrm{T}({\sim}A)$ will be true at the limit. It follows that S is stably true.

In conclusion, we have the stability of the following logic.

The logic ST

DIS $\mathrm{T}(A) \rightarrow \mathrm{C}(A)$
IP $\mathrm{T}(A \rightarrow B) \rightarrow (\mathrm{T}(A) \rightarrow \mathrm{T}(B))$
BAR $\forall x \mathrm{T}(A) \rightarrow \mathrm{T}(\forall x A)$
S $\mathrm{T}(\mathrm{T}(A) \rightarrow A) \rightarrow \mathrm{P}(A)$
NEC′ If $\mathrm{ST} \vdash A$ then $\mathrm{ST} \vdash \mathrm{T}(A)$.

This logic is worthy of further investigation. We shall explore it further on another occasion.

Next consider the **S4** axiom which takes the form:

S4 $\mathrm{T}(A) \rightarrow \mathrm{T}(\mathrm{T}(A))$.

Theorem 5.4.4 S4 *is sound.*

Proof: S4 insists that if A is stably true then it is stably true that it is stably true. To see that this is so one only has to observe that by the rule of revision at successor ordinals the positive stability of A implies the positive stability of $\mathrm{T}(A)$.

Once again, although S4 is sound it is not stably true. To see this we first observe that the following axiom is sound:

IPT $T(T(A) \rightarrow T(B)) \rightarrow T(A \rightarrow B)$.

Theorem 5.4.5 IPT *is sound.*

Proof: This is obvious given the definition of truth at successor ordinals.

IPT is not stably true: if it were then applying IPT to the assertion of its stable truth would yield $T((T(A) \rightarrow T(B)) \rightarrow (A \rightarrow B))$, which by axiom T gives $(T(A) \rightarrow T(B)) \rightarrow (A \rightarrow B)$. Any two sentences A, B where A is true but not stably true and B is false provide a counterexample to this principle.

We can now show that S4 cannot be stably true. Assume that it is, i.e., $T(T(A) \rightarrow T(T(A)))$, then by IPT we obtain $T(A \rightarrow T(A))$, which by axiom T yields $A \rightarrow T(A)$, and this together with T yields $A \leftrightarrow T(A)$, the Tarski biconditionals.

We now consider the **S5** axiom. The addition of the **S5** axiom $C(A) \rightarrow T(C(A))$ renders the theory inconsistent: from S5 and T we then obtain $\sim T(\sim A) \leftrightarrow T(\sim T(\sim A))$.

Unfortunately, the Liar sentence satisfies $T(B \leftrightarrow T(\sim B))$, and so we obtain (writing B for A in the above) the following equivalences: $\sim T(\sim B) \leftrightarrow T(\sim T(\sim B)) \leftrightarrow T(\sim B)$. Indeed, in terms of stability S5 is clearly false since it insists that any assertion which is not stably false is stably, not stably false.

From axioms T and S4 we can deduce that $T(A) \leftrightarrow T(T(A))$, so if A is stably true it is stably true that it is stably true. We cannot deduce that A is stably false iff $T(A)$ is stably false, but we could if we added the following: $\sim T(A) \rightarrow T(\sim T(A))$. But a special instance of this is the already discarded S5 axiom. Indeed, $\sim T(A) \rightarrow T(\sim T(A))$ in conjunction with M allows the derivation of all instances of the Tarski biconditionals. However, we do have the following implication already:

R $T(A) \rightarrow T(C(A))$.

This follows from SDIS, IMP, T, and S4. Surprisingly the converse of R is also sound:

L $T(C(A)) \rightarrow T(A)$.

Theorem 5.4.6 L *is sound.*

Proof: If $\sim T(A)$ is true from some ordinal, then by the definition of revision at successor ordinals $\sim A$ will be true from some ordinal onwards.

We often refer to the conjunction of R and L as NEG. Indeed, L is a special case of IPT. Neither of the stable analogues of R and L is sound; indeed, either statement leads to a contradiction. The soundness of IPT destroys the possibility: the stable truth of R leads by IPT to $T(T(A) \rightarrow A)$ and the stable truth of L leads by IPT to $T(A \rightarrow T(A))$, both of which are unacceptable.

Let LS be the logic D + T + S + S4 + IPT + NEG.

Theorem 5.4.7 *In* LS *we have:*
(i) $P(A) \leftrightarrow P(T(A))$
(ii) $P(A) \leftrightarrow T(P(A))$

(iii) $P(A) \leftrightarrow (T(A) \leftrightarrow A \;\&\; F(A) \leftrightarrow \sim A)$
(iv) $P(A) \leftrightarrow T(T(A) \rightarrow A)$.

Proof: For part (i) we use S4, T, and NEG. For the second part we employ T and S4. Part (iii) employs M and T. For (iv) observe that from $T(A)$, the tautology $A \rightarrow (T(A) \rightarrow A)$, NEC, and IP we obtain $T(T(A) \rightarrow A)$, and from $T(\sim A)$, R, the tautology $\sim B \rightarrow (B \rightarrow C)$, NEC, and IP we again obtain $T(T(A) \rightarrow A)$.

From (iv) and S we see that $T(T(A) \rightarrow A)$ is both necessary and sufficient for stability and that the positive and negative statements of the Tarski biconditionals are also necessary and sufficient.

With the T, S4, and NEG axioms we can also derive inference rules for iterated truth:

$$\frac{T(A)}{T(T(A))} \qquad \frac{T(T(A))}{T(A)}$$

$$\frac{T(\sim A)}{T(\sim T(A))} \qquad \frac{T(\sim T(A))}{T(\sim A)}$$

The lack of a full axiom of Necessitation means that substitution fails in LS. However, a weaker principle is derivable. First define:

$$A \simeq B =_{def} T(A \leftrightarrow B).$$

Then the following principle:

$\simeq$**Sub** $A \simeq B \rightarrow \Psi[A] \simeq \Psi[B]$

is derivable, where Ψ is any context in which a wff can be meaningfully substituted.

Theorem 5.4.8 *The above principle is derivable in* LS.

Proof: Use induction on the context. The case where the context is T( ) is taken care of by SIP.

This completes our discussion of the logic of stable truth. There is no doubt a great deal more which could be said in regard to its location within the spectrum of modal logics, but we hope to have done enough to convince the reader that the logic of stable truth constitutes a quite rich modal theory of a rather distinctive kind. For those who think that all is sweet we end on a negative note.

Theorem 5.4.9 *In the logic of stable truth the notion of truth is not* internally definable.

Proof: By this we mean that $\forall z P(T(z))$ is provably false. Suppose not. Let A be the Liar, i.e., $A \leftrightarrow T(\sim A)$, then we have $T(T(A)) \vee F(T(A))$. The first disjunct leads via axiom T to $T(A)$ & A which yields a contradiction, while the second leads by NEG and T to $T(\sim A)$ & $\sim A$, again a contradiction.

This is the stability version of the result of Aczel [1].

REFERENCES

[1] Aczel, P., "Frege structures and the notions of proposition, truth and set," pp. 31–39 in *The Kleene Symposium*, edited by J. Barwise, H. J. Keisler, and K. Kunan, North Holland Studies in Logic 101, North Holland, Amsterdam, 1980.

[2] Barendregt, H., *The Lambda Calculus: Its Syntax and Semantics*, North Holland Studies in Logic 103, North Holland, Amsterdam, 1984.

[3] Feferman, S., "Towards useful type-free theories, I," *The Journal of Symbolic Logic*, vol. 49 (1984), pp. 75–111.

[4] Gilmore, P. C., "The consistency of partial set theory without extensionality," in *Proceedings of Symposia in Pure Mathematics*, vol. 13, part II, American Mathematical Society, 1974.

[5] Gupta, A., "Truth and paradox," *Journal of Philosophical Logic*, vol. 11 (1982), pp. 1–60.

[6] Herzberger, H., "Notes on naive semantics," *Journal of Philosophical Logic*, vol. 11 (1982), pp. 61–102.

[7] Hindley, R. and J. Seldin, *Introduction to Combinatory Logic*, Cambridge University Press, Cambridge, 1986.

[8] Kaplan, D. and R. Montague, "A paradox regained," *Notre Dame Journal of Formal Logic*, vol. 1 (1960), pp. 79–90.

[9] Kripke, S., "Outline of a theory of truth," *The Journal of Philosophy*, vol. 72 (1975), pp. 690–716.

[10] Scott, D., "Combinators and classes," pp. 1–26 in *Lambda Calculus and Computer Science*, Lecture Notes in Computer Science 37, Springer-Verlag, Berlin, 1975.

[11] Turner, R., "A theory of properties," *The Journal of Symbolic Logic*, vol. 52 (1987), pp. 445–472.

Department of Computer Science
Wivenhoe Park
University of Essex
Colchester C04 3SQ, England