TRUTH AND INDUCTIVE DEFINABILITY

RICHARD G. HECK

The purpose of this note is to give an exposition of Kripke's theory of truth [4] within the context of the theory of induction on abstract structures. The results are not presented in full generality: That is not needed for the applications made here. For that, see Moschovakis [6], Kechris and Moschovakis, [2], and McGee [5, ch. 5]. Useful papers here, besides Kripke's original, include Feferman [1]. A background in basic set theory, together with some mathematical sophistication, should be enough for the reader to understand this note.

1. ORDINALS

We shall need some basic facts about ordinal numbers.[1] Intuitively, an *ordinal number* answers the question which position in a sequence is occupied by a given element. The sequences in question must be *non-repeating*—no element may occur more than once—and *well-ordered*. A sequence is well-ordered if it is linearly ordered—transitive, irreflexive, and totally ordered—and the sequence must satisfy an analogue of the least number principle: Given any non-empty set S of elements of the sequence, there must be a least one, that is, a member of S that comes before every other member of S. If we think of the sequence as ordered by the relation $<$, then the conditions just mentioned may be formalized as:

$$\forall x \forall y \forall z (x < y \wedge y < z \rightarrow x < z)$$
$$\forall x \forall y (x < y \rightarrow \neg y < x)$$
$$\forall x \in S \forall y \in S (x < y \vee x = y \vee y < x)$$
$$\forall x (x \subseteq S \rightarrow \exists y \in x \forall z \in x (y \leq z)]$$

The natural numbers 1, 2, 3, and so forth can be construed as ordinals, pronounced 'first', 'second', 'third', and so forth.

There are also infinite ordinals. Consider the following sequence:

$$1, 2, 3, \ldots, \text{Julius Caesar}, \text{Brutus}, \text{Claudius}$$

This is clearly a non-repeating, well-ordered sequence. Each natural number n occupies the n^{th} position in the sequence. What position does Caesar occupy? This position is called the ω^{th} position: The first infinite ordinal is ω (omega). The next position, occupied by Brutus, is the $(\omega+1)^{\text{st}}$ position; the next, occupied by Claudius, the $(\omega + 2)^{\text{nd}}$. And so on.[2]

Consider now the following sequence:

$$1, 2, 3, \ldots, -1, -2, -3, \ldots, \text{Caesar}, \text{ Brutus}, \text{ Claudius}.$$

[1] The theory of ordinal numbers can be developed in set theory, which treats ordinal numbers as sets of a certain kind. See any decent textbook.

[2] Note, by the way, that ordinal addition is not commutative: $2 + \omega$ is just ω again, since the thing that occupies the ω^{th} place after the 2^{nd} thing is just the ω^{th} thing.

Again, each natural number n occupies the n^{th} position of the sequence; each negative integer $-n$ occupies the $(\omega + n)^{\text{th}}$ position. What position does Caesar occupy here? This is the $(\omega+\omega)^{\text{th}}$ or $(2\omega)^{\text{th}}$ position; Brutus occupies the $(2\omega+1)^{\text{st}}$; Claudius, the $(2\omega + 2)^{\text{nd}}$. And so on, through 3ω, 4ω, ...,$n\omega$, and onward, until we reach $\omega \times \omega$ or ω^2, which is the position Caesar occupies in this sequence:

$$< 1, 1 >, < 1, 2 >, \ldots, < 2, 1 >, < 2, 2 >, \cdots < n, 1 >, < n, 2 >, \ldots, \text{Caesar}.$$

And so on through $\omega^2 + \omega$, $\omega^2 + 2\omega$,$\omega^2 + \omega^2 = 2(\omega^2)$,$3(\omega^2)$,$\omega(\omega^2) = \omega^3$, ω^4, ω^ω, and so on (and on) for a very long time.

One can see here that there are two kinds of ordinals: There are *successor* ordinals and there are *limit* ordinals. We say that α is a successor ordinal if $\alpha = \beta + 1$, for some ordinal β. Otherwise, α is a limit. Examples of limit ordinals are 0 (which is the only finite limit),[3] ω, and ω^2.

All of the ordinals mentioned so far are *countable*, meaning that there are only as many objects in any such sequence as there are natural numbers: Indeed, any such sequence can be constructed simply from natural numbers; for example, the elements $< m, k >$ of the last displayed sequence may be replaced by $2^m 3^k$; Caesar may be replaced by 5.

The countable ordinals themselves, under the obvious ordering, form a non-repeating, well-ordered sequence. If we add Caesar to the end of this sequence, we may ask what position he occupies. The answer is ω_1, the first *uncountable* ordinal. And of course the ordinals do not stop there. But we shall not need to go further. The one fact we do need is that the set of countable ordinals is not itself a countable set. More precisely, what we need to know is just this.

Fact 1. *Let C be the set of all countable ordinals. Suppose that f is a one-one function from C into (not necessarily onto) some set S. Then S is uncountable.*

For a proof, see any decent textbook on set theory.

The other fact we shall need about the ordinals is that, just as one can show that all natural numbers have some property by means of mathematical induction, it is possible to show that all ordinals have some property by means of *transfinite* induction. In the finite case, one typically argues as follows: One shows that 0 is F—the basis case—and one then shows that, if a given number n is F, so is $n + 1$—the induction step. An equivalent method of proof replaces that sort of induction step with this one: If all numbers less than n are F, then so is n.[4] One uses this kind of induction, for example, in proving that every number has a unique prime factorization: One does not simply assume that n has a unique prime factorization and then show that $n + 1$ does; rather, one assumes that *every* number less than n has a unique prime factorization and then shows that n does.

Transfinite induction is simply this same method of proof, but applied to all ordinals (or all ordinals less than some given ordinal, e.g., to all countable ordinals). That is, to show that all ordinals are F, one shows that $F0$ and that, if $\forall\beta < \alpha(F\beta) \rightarrow F\alpha$. The induction step often divides into different cases for successor ordinals and limit ordinals.

[3] Strictly speaking, there is no ordinal 0, since there is no zero$^{\text{th}}$ position of a sequence. But we allow 0 for convenience.

[4] The latter follows from the former: Suppose we want to show that $\forall x Fx$ and we know that $\forall n[\forall x < n(Fx) \rightarrow Fn]$. Then consider $\forall x < n(Fx)$ and argue by induction in the usual way.

Similarly, in the finite case, we can define functions on the natural numbers by recursion: One says what $\phi(0)$ is supposed to be and then, assuming one knows what $\phi(n)$ is, one says what $\phi(n+1)$ is to be. Again, there is a similar but somewhat different method: Assuming one knows what $\phi(k)$ is, for all $k < n$, one says what $\phi(n)$ is to be. Functions on the ordinals can be defined by this method, as well, in which case it is known as *transfinite recursion*. We'll see examples below.

2. Monotone Operators

Let A be a set. A^n is the n^{th} Cartesian product of A with itself, that is, the set of sequences $<a_1, \ldots, a_n>$ with elements from A. It is easy to show that A^n is countable if A is. The *power set* of A, $\wp(A)$ is the set of all subsets of A. If A is countably infinite, then $\wp(A)$ is *un*countably infinite, by a simple diagonalization due to Cantor.

Let A be a set of sets. We say that an operator $\Phi : A \to A$ is *monotonic* if, whenever $S \subseteq T$, $\Phi(S) \subseteq \Phi(T)$. Our attention will be focused upon operators $\Phi : \wp(A^n) \to \wp(A^n)$ that map sets of n-tuples whose elements are taken from A to sets of n-tuples whose elements are taken from A.

The key theorem is that monotonic operators always have fixed points, in the following sense:

Definition. S is a *fixed point* of Φ if $\Phi(S) = S$.

The idea behind the proof is simple. We start with $\emptyset$ and start applying Φ, forming the sequence $\emptyset \subseteq \Phi(\emptyset) \subseteq \Phi(\Phi(\emptyset)) \subseteq \cdots$ and continue in this way. The various inclusions hold because Φ is monotone. Obviously, the first one must hold; but then the second one holds; and so forth. If we don't hit a fixed point before we run out of finite ordinals, we take the union of everything we've had so far and continue applying Φ, taking unions at any limits we reach. Eventually, we will reach a fixed point, for reasons we shall see below.

We now give a real proof of this result.

Definition. If Φ is monotonic, then the function I^x_Φ is defined by transfinite recursion thus:

$$I^0_\Phi = \emptyset$$
$$I^\alpha_\Phi = \Phi(\bigcup_{\xi<\alpha} I^\xi_\Phi), \text{ for } \alpha \neq 0$$

Theorem 2 (Fixed-point Theorem)**.** *Let $\Phi : A \to A$ be monotonic. Then Φ has a fixed point. Moreover, if A is countable, then there is a countable ordinal α such that $\Phi(I^\alpha_\Phi) = I^\alpha_\Phi$, and I^α_Φ is the minimal fixed point of Φ, in the sense that it is a subset of every other fixed point of Φ.*[5]

Lemma 3. *If Φ is monotonic, then $\forall\beta\forall\alpha(\beta < \alpha \to I^\beta_\Phi \subseteq I^\alpha_\Phi)$. In particular, $I^\alpha_\Phi \subseteq I^{\alpha+1}_\Phi$.*

Proof. By transfinite induction on α. This holds vacuously for $\alpha = 0$. So suppose $\alpha \neq 0$. We want to see that $\forall\beta < \alpha(I^\beta_\Phi \subseteq I^\alpha_\Phi)$. So let $\beta < \alpha$. Then obviously $\bigcup_{\xi<\beta} I^\xi_\Phi \subseteq \bigcup_{\xi<\alpha} I^\xi_\Phi$, since the latter union includes all the sets the former one

[5] More generally, the α in question will be less than $|\, A \,|^+$, which is the least cardinal greater than the cardinality of A.

does. So, since Φ is monotone, we have $I^{\beta}_{\Phi} = \Phi(\bigcup_{\xi<\beta} I^{\xi}_{\Phi}) \subseteq \Phi(\bigcup_{\xi<\alpha} I^{\xi}_{\Phi}) = I^{\alpha}_{\Phi}$, as claimed. □

Corollary 4. $I^{\alpha+1}_{\Phi} = \Phi(I^{\alpha}_{\Phi})$

Proof. By definition, $I^{\alpha+1}_{\Phi} = \Phi(\bigcup_{\xi<\alpha+1} I^{\xi}_{\Phi})$. I claim that $\bigcup_{\xi<\alpha+1} I^{\xi}_{\Phi} = I^{\alpha}_{\Phi}$. Now, $\bigcup_{\xi<\alpha+1} I^{\xi}_{\Phi} = I^{\alpha}_{\Phi} \cup \bigcup_{\xi<\alpha} I^{\xi}_{\Phi}$. But if $\xi < \alpha$, then $I^{\xi}_{\Phi} \subseteq I^{\alpha}_{\Phi}$, so $\bigcup_{\xi<\alpha} I^{\xi}_{\Phi} \subseteq I^{\alpha}_{\Phi}$. Hence, $I^{\alpha}_{\Phi} \cup \bigcup_{\xi<\alpha} I^{\xi}_{\Phi} = I^{\alpha}_{\Phi}$. □

Proof of Theorem 2. Suppose that for no countable α do we have $\Phi(I^{\alpha}_{\Phi}) = I^{\alpha}_{\Phi}$. Then for every α, $I^{\alpha}_{\Phi} \subsetneq \Phi(I^{\alpha}_{\Phi}) = I^{\alpha+1}_{\Phi}$. Hence, for each countable α, we can find $\varphi(\alpha) \in I^{\alpha+1}_{\Phi} \setminus I^{\alpha}_{\Phi}$.[6] Now suppose $\alpha \neq \beta$; either $\alpha < \beta$ or conversely; without loss of generality, suppose $\alpha < \beta$. Then $\phi(\beta) \in I^{\beta+1}_{\Phi} \setminus I^{\beta}_{\Phi}$, whence $\phi(\beta) \notin I^{\beta}_{\Phi}$, whereas $\phi(\alpha) \in I^{\alpha+1}_{\Phi}$. And since $\alpha < \beta$, $\alpha + 1 \leq \beta$, whence $I^{\alpha+1}_{\Phi} \subseteq I^{\beta}_{\Phi}$, and so $\varphi(\alpha) \neq \phi(\beta)$. But that means that φ is one-to-one, whence it is a function from all the countable ordinals into A, which is impossible.

So there is a countable ordinal α such that $\Phi(I^{\alpha}_{\Phi}) = I^{\alpha}_{\Phi}$. Henceforth let α be the least such ordinal. So I^{α}_{Φ} is a fixed point of Φ.

Let F be any fixed point of Φ. We show by transfinite induction that $I^{\beta}_{\Phi} \subseteq F$, for all β, from which it follows that $I^{\alpha}_{\Phi} \subseteq F$. Obviously, if $\beta = 0$, then $I^{0}_{\Phi} = \emptyset \subseteq F$. So suppose $I^{\xi}_{\Phi} \subseteq F$, for all $\xi < \beta$. Then $\bigcup_{\xi<\beta} I^{\xi}_{\Phi} \subseteq F$,[7] so $I^{\beta}_{\Phi} = \Phi(\bigcup_{\xi<\beta} I^{\xi}_{\Phi}) \subseteq \Phi(F) = F$, by monotonicity. □

Notation. We write: I_{Φ}, for the minimal fixed point of Φ.

3. Inductive Definitions

Definition. A predicate letter R is said to *occur positively* in a formula ϕ if, and only if, it is in the smallest class C of formulae containing:

(1) All formulae in which R does not occur;
(2) $R(t_1, \ldots, t_n)$, for any terms t_i;
(3) the conjunctions, disjunctions, and existential or universal quantifications of any formulae in C.

What this means, in essence, is that R does not occur inside negation or in the *antecedent* of a conditional (since $p \to q$ can be defined as: $\neg p \vee q$). We call a formula that is logically equivalent to a formula in which R occurs positively *R-positive*.

Let $\mathcal{L}$ be a language; $\mathcal{T}$, a theory in $\mathcal{L}$. We will be interested in languages $\mathcal{L}^*$ that result from adding to $\mathcal{L}$ some new relation symbols of various arities. *Par abuse de langue*, we shall also think of $\mathcal{T}$ as a theory in $\mathcal{L}^*$, with the same axioms. We shall write these new relation constants as lower case Greek letters, σ, τ, and the like, and indicate their arity with superscripts as necessary.

Our main theorem is this:

Theorem 5 (Inductive Definitions)**.** *Let $\mathcal{T}$ be a theory in $\mathcal{L}$; let $\mathcal{L}^*$ be $\mathcal{L}$ plus a new n-place relation symbol σ^n; and let $\phi(x_1, \ldots, x_n, \sigma^n)$ be a formula of $\mathcal{L}^*$ in*

[6] The Axiom of Choice is needed for this step.

[7] If $\forall x \in S(x \subseteq T)$, then $\cup S \subseteq T$.

which σ^n occurs positively and in which only the variables shown are free. Then the theory $\mathcal{T}^*$, which is the result of adding the new axiom

(σ) $$\forall x_1 \ldots \forall x_n[\sigma^n(x_1, \ldots, x_n) \equiv \phi(x_1, \ldots, x_n, \sigma^n)]$$

to $\mathcal{T}$, is a conservative extension of $\mathcal{T}$. A fortiori, $\mathcal{T}^*$ is consistent if $\mathcal{T}$ is consistent.

Definition. We say that σ^n is *inductively defined* by the formula $\phi(x_1, \ldots, x_n, \sigma^n)$.

Before we give the proof, a couple examples. Let $\mathcal{L}$ be the language of arithmetic, and let $\mathcal{T}$ be PA. First, we can inductively define 2^x. Let $\psi(x, y, \mathbf{exp})$ be:

$$(x = 0 \wedge y = 1) \vee (x = n \wedge \exists z(\mathbf{exp}(n, z) \wedge y = 2 \times z)$$

Then exp occurs positively in ψ so, by the theorem, the result of adding the axiom:

$$\mathbf{exp}(x, y) \equiv (x = 0 \wedge y = 1) \vee (x = n \wedge \exists z(\mathbf{exp}(n, z) \wedge y = 2 \times z)$$

to PA is consistent. This, as said, gives an inductive characterization of 2^x.

Second, we can give an inductive definition of the notion of a *term* of the language. Assume that we have already introduced the basics of Gödel numbering into the system. Let $\phi(x, \tau)$ be the formula:

$$\begin{aligned} x = {}& \ulcorner 0 \urcorner \vee \\ \exists y(x = {}& \ulcorner Sy \urcorner \wedge \tau(y)) \vee \\ \exists z \exists y(x = {}& \ulcorner y + z \urcorner \wedge \tau(y) \wedge \tau(z)) \vee \\ \exists z \exists y(x = {}& \ulcorner y \times z \urcorner \wedge \tau(y) \wedge \tau(z)) \end{aligned}$$

Then τ occurs positively in ϕ and so, by the theorem, the result of adding the axiom:

(Def τ) $$\begin{aligned} \forall x[\tau(x) \equiv \quad & x = \ulcorner 0 \urcorner \vee \\ & \exists y(x = \ulcorner Sy \urcorner \wedge \tau(y)) \vee \\ & \exists z \exists y(x = \ulcorner y + z \urcorner \wedge \tau(y) \wedge \tau(z)) \vee \\ & \exists z \exists y(x = \ulcorner y \times z \urcorner \wedge \tau(y) \wedge \tau(z))] \end{aligned}$$

to PA is a conservative extension of PA. This axiom characterizes the set of terms inductively.

Before we prove Theorem 5, we shall first need to establish some properties of positive formulae. Let an interpretation $\mathcal{M}$ of $\mathcal{L}$ be given, with domain $\mathcal{D}$. Let S be a subset of $\mathcal{D}^n$—i.e., a possible extension for σ^n—and let $\mathcal{M}_S$ be the expansion of $\mathcal{M}$ to an interpretation of $\mathcal{L}^*$ that we get by assigning S to σ^n. Now suppose we begin with an interpretation $\mathcal{M}_S$ of $\mathcal{L}^*$. In this interpretation, some sequences will satisfy $\phi(x_1, \ldots, x_n, \sigma^n)$ and some will not. The set of sequences that do satisfy it is *another* possible extension for σ^n. So there is a natural operator $\Sigma_\phi^{\mathcal{M}}$ on $\mathcal{D}^n$ that we can define as follows:

$$\Sigma_\phi^{\mathcal{M}}(S) = \{\langle y_1, \ldots, y_n\rangle : \langle y_1, \ldots, y_n\rangle \text{ satisfies } \phi(x_1, \ldots, x_n, \sigma^n) \text{ in } \mathcal{M}_S\}$$

Thus, $\Sigma_\phi^{\mathcal{M}}(S)$ is the extension of ϕ in $\mathcal{M}_S$.

Now for the key fact:

Lemma 6. *If $\phi(x_1, \ldots, x_n, \sigma^n)$ is σ^n-positive, then $\Sigma_\phi^{\mathcal{M}}$ is monotonic and so has a minimal fixed point $I_\phi^{\mathcal{M}}$.*

Proof. The monotonicity of $\Sigma_\phi^{\mathcal{M}}$ will follow immediately from the next result. The existence of a fixed pont then follows from Theorem 2. □

Proposition 7. *Let $\phi(x_1, \ldots, x_n, \sigma^n)$ be a σ^n-positive formula in $\mathcal{L}^*$, and let $\mathcal{M}$ be an interpretation of $\mathcal{L}$ with domain $\mathcal{D}$. Suppose that $S \subseteq T \subseteq \mathcal{D}^n$. Then if a sequence s satisfies $\phi(x_1, \ldots, x_n, \sigma^n)$ in $\mathcal{M}_S$, s also satisfies $\phi(x_1, \ldots, x_n, \sigma^n)$ in $\mathcal{M}_T$.*

Proof. The proof is by induction on the complexity of σ^n-positive formulas.

The basis case concerns formulas in which σ^n does not occur and atomic formulae of the form $\sigma^n(x_{k_1}, \ldots, x_{k_n})$, for some $k_1, \ldots, k_n$. In the former case, the change from $\mathcal{M}_S$ to $\mathcal{M}_T$ is irrelevant. In the latter, if s satisfies $\sigma^n(x_{k_1}, \ldots, x_{k_n})$ in $\mathcal{M}_S$, then $<s(k_1), \ldots, s(k_n)> \in S \subseteq T$, whence s satisfies $\sigma^n(x_{k_1}, \ldots, x_{k_n})$ in $\mathcal{M}_T$.

So suppose the proposition holds for A and B; we want to show it holds for their conjunction and disjunction. But if s satisfies $A \wedge B$ in $\mathcal{M}_S$, then it satisfies both A and B in $\mathcal{M}_S$; hence, by the induction hypothesis, it satisfies both A and B in $\mathcal{M}_T$; so it satisfies $A \wedge B$ in $\mathcal{M}_T$. The argument for disjunction is similar.

So suppose the proposition holds for $A(x_i)$, where x_i occurs free in $A(x_i)$, as may additional variables. We want to show that the proposition holds for $\exists x_i A(x_i)$ and $\forall x_i A(x_i)$. If s satisfies $\exists x_i A(x_i)$ in $\mathcal{M}_S$, then there is a sequence t such that $\forall j \neq i(t(j) = s(j))$ and t satisfies $A(x_i)$ in $\mathcal{M}_S$. By the induction hypothesis, t satisfies $A(x_i)$ in $\mathcal{M}_T$; but then s satisfies $\exists x_i A(x_i)$ in $\mathcal{M}_T$. The argument for the universal quantifier is similar. □

Our goal, recall, is to show that the result $\mathcal{T}^*$ of adding the new axiom (σ) to $\mathcal{T}$ is a conservative extension of $\mathcal{T}$. To do so, we show that any model $\mathcal{M}$ of $\mathcal{T}$ can be expanded to a model of $\mathcal{T}^*$, that is, that we can expand $\mathcal{M}$ in such a way as to make (σ) true.[8] The model in question will be the one that assigns $I_\phi^{\mathcal{M}}$ to σ^n. That this does the trick is the content of the following.

Lemma 8. *Let $\mathcal{M}$ be an interpretation of $\mathcal{L}$; let $\phi(x_1, \ldots, x_n, \sigma^n)$ be a σ^n-positive formula in $\mathcal{L}^*$; and let I be any fixed point of $\Sigma_\phi^{\mathcal{M}}$. Then*

$$(\sigma) \qquad \forall x_1 \ldots \forall x_n [\sigma^n(x_1, \ldots, x_n) \equiv \phi(x_1, \ldots, x_n, \sigma^n)]$$

is true in $\mathcal{M}_I$.

Proof. The formula mentioned is true in $\mathcal{M}_I$ iff

$$\{< y_1, \ldots, y_n >:< y_1, \ldots, y_n > \text{ satisfies } \sigma^n(x_1, \ldots, x_n) \text{ in } \mathcal{M}_I\} =$$
$$\{< y_1, \ldots, y_n >:< y_1, \ldots, y_n > \text{ satisfies } \phi(x_1, \ldots, x_n, \sigma^n) \text{ in } \mathcal{M}_I\}$$

But

$$\{< y_1, \ldots, y_n >:< y_1, \ldots, y_n > \text{ satisfies } \sigma^n(x_1, \ldots, x_n) \text{ in } \mathcal{M}_I\} = I,$$

by the definition of $\mathcal{M}_I$;

$$I = \Sigma_\phi^{\mathcal{M}}(I),$$

since I is a fixed point of $\Sigma_\phi^{\mathcal{M}}$; and

$$\Sigma_\phi^{\mathcal{M}}(I) = \{< y_1, \ldots, y_n >:< y_1, \ldots, y_n > \text{ satisfies } \phi(x_1, \ldots, x_n, \sigma^n) \text{ in } \mathcal{M}_I\},$$

by the definition of $\Sigma_\phi^{\mathcal{M}}$. □

We can now complete the proof of theorem 5.

[8] Thus, $\mathcal{T}^*$ is, as it is said, *semantically* conservative over $\mathcal{T}$.

Proof of Theorem 5. Let $\mathcal{M}$ be any model of $\mathcal{T}$. We shall show that $\mathcal{M}_{I_\phi^{\mathcal{M}}}$, which is an expansion of $\mathcal{M}$, is a model of $\mathcal{T}^*$. All axioms of $\mathcal{T}$ continue to be true in $\mathcal{M}_{I_\phi^{\mathcal{M}}}$, since we have not changed the interpretation of $\mathcal{L}$. And by lemma 8, (σ) is also true in $\mathcal{M}_{I_\phi^{\mathcal{M}}}$. So we are done. □

A simple generalization of the above arguments yields the following.

Theorem 9 (Simultaneous Inductive Definitions)**.** *Let $\mathcal{T}$ be a theory in $\mathcal{L}$; let $\psi(y_1, \ldots, y_m, \sigma^m, \tau^n)$ and $\phi(x_1, \ldots, x_n, \sigma^m, \tau^n)$ be formulae of $\mathcal{L}^*$ in which both σ^m and τ^n occur positively. Then the theory $\mathcal{T}^*$ which is the result of adding the new axioms*

(σ) $\qquad \forall y_1 \ldots \forall y_m[\sigma^m(y_1, \ldots, y_m) \equiv \psi(y_1, \ldots, y_m, \sigma^m, \tau^n)$

(τ) $\qquad \forall x_1 \ldots \forall x_n[\tau^n(x_1, \ldots, x_n) \equiv \phi(x_1, \ldots, x_n, \sigma^m, \tau^n)$

to $\mathcal{T}$, is a conservative extension of $\mathcal{T}$. A fortiori, $\mathcal{T}^*$ is consistent if $\mathcal{T}$ is consistent.

Proof. Given an interpretation $\mathcal{M}$ of $\mathcal{L}$, let $\mathcal{M}_{S,T}$ be the expansion of $\mathcal{M}$ to an interpretation of $\mathcal{L}^*$ in which S is the extension of σ and T is the extension of τ. Define an operator $\Sigma_\psi^{\mathcal{M}}(S,T)$ as follows: $\Sigma_\psi^{\mathcal{M}}(S,T)$ is the extension of σ^m in $\mathcal{M}_{S,T}$; that is:

$$\Sigma_\psi^{\mathcal{M}} = \{<y_1, \ldots, y_m> : y_1, \ldots, y_m \text{ satisfies } \psi(y_1, \ldots, y_m, \sigma^m, \tau^n) \text{ in } \mathcal{M}_{S,T}\}$$

Define $\Sigma_\phi^{\mathcal{M}}(S,T)$ similarly.

By simultaneous induction, define:

$$\Psi_0 = \Sigma_\psi^{\mathcal{M}}(\emptyset, \emptyset)$$
$$\Phi_0 = \Sigma_\phi^{\mathcal{M}}(\emptyset, \emptyset)$$
$$\Psi_\alpha = \Sigma_\psi^{\mathcal{M}}(\bigcup_{\xi<\alpha} \Psi_\xi, \bigcup_{\xi<\alpha} \Phi_\xi)$$
$$\Phi_\alpha = \Sigma_\phi^{\mathcal{M}}(\bigcup_{\xi<\alpha} \Psi_\xi, \bigcup_{\xi<\alpha} \Phi_\xi)$$

We can then show, just as in the proof of Theorem 5, that $\Sigma_\psi^{\mathcal{M}}$ and $\Sigma_\phi^{\mathcal{M}}$ have *simultaneous* fixed points $I_\psi^{\mathcal{M}}$ and $I_\phi^{\mathcal{M}}$: That is, $I_\psi^{\mathcal{M}} = \Sigma_\psi^{\mathcal{M}}(I_\psi^{\mathcal{M}}, I_\phi^{\mathcal{M}})$ and $I_\phi^{\mathcal{M}} = \Sigma_\phi^{\mathcal{M}}(I_\psi^{\mathcal{M}}, I_\phi^{\mathcal{M}})$. And we can then show that $\mathcal{M}_{I_\psi^{\mathcal{M}}, I_\phi^{\mathcal{M}}}$ is a model of $\mathcal{T}^*$, again as before. □

Obviously, this can be extended to any finite number of formulas.

Now let $\mathcal{L}_{\mathcal{M}}$ be an *interpreted* language, consisting of a language $\mathcal{L}$ and an interpretation $\mathcal{M}$ for $\mathcal{L}$ with domain $\mathcal{D}$; the intepretation is thus fixed. Suppose that $\phi(x_1, \ldots, x_n, \sigma^n)$ is a σ^n-positive formula in $\mathcal{L}^*_{\mathcal{M}}$. It follows from Theorem 2 that the operator $\Sigma_\phi^{\mathcal{M}}$ has a fixed point $I_\phi^{\mathcal{M}}$, which we may just call 'I_ϕ'. And it then follows from Lemma 8 that

(σ) $\qquad \forall x_1 \ldots \forall x_n[\sigma^n(x_1, \ldots, x_n) \equiv \phi(x_1, \ldots, x_n, \sigma^n)]$

is true when σ^n is assigned I_ϕ as its extension.

Definition. If a set S is I_ϕ for some appropriate formula ϕ of $\mathcal{L}^*_{\mathcal{M}}$, then S is said to be a *fixed point over $\mathcal{L}_{\mathcal{M}}$*.

Definition. Let $R \subseteq D^m$. We say that R is *inductive* over $\mathcal{L}_\mathcal{M}$ if there is a fixed point I_ϕ over $\mathcal{L}_\mathcal{M}$ and there are objects $a_1, \ldots, a_{n-m}$ such that:

$$< x_1, \ldots, x_m > \in R \equiv < a_1, \ldots, a_{n-m}, x_1, \ldots, x_m > \in I_\phi.$$

In effect, a relation R is inductive just in case it is a 'planar section' of a fixed point or, again, if it is 'parametrically definable' in terms of a fixed point.

Definition. A relation whose complement is inductive is called *co-inductive*; a relation that is both inductive and co-inductive is said to be *hyperelementary*. If S is actually defined by a formula $\phi(x_1, \ldots, x_n)$ of $\mathcal{L}_\mathcal{M}$ itself, we say that it is *elementary*.

Here are a few basic facts about inductive relations, which we record without proof.

Proposition 10.

(1) *Every relation elementary in $\mathcal{L}_\mathcal{M}$ is inductive over $\mathcal{L}_\mathcal{M}$.*
(2) *Relations inductive over $\mathcal{L}_\mathcal{M}$ are closed under union and intersection.*
(3) *Relations hyperelementary over $\mathcal{L}_\mathcal{M}$ are closed under complementation.*
(4) *The union and intersection of two inductive relations are inductive.*

Lemma 11 (Simultaneous Induction Lemma)**.** *Let $\psi(y_1, \ldots, y_m, \sigma^m, \tau^n)$ and $\phi(x_1, \ldots, x_n, \sigma^m, \tau^n)$ be formulae of $\mathcal{L}^*_\mathcal{M}$ in which both σ^m and τ^n occur positively. Let $\Sigma_{\mathcal{M},\psi}(S,T)$ and $\Sigma_{\mathcal{M},\phi}(S,T)$ be the operators naturally associated with these formulae, as above. Then their minimal simultaneous fixed points $I_{\mathcal{M},\psi}$ and $I_{\mathcal{M},\phi}$ are both inductive over $\mathcal{L}_\mathcal{M}$.*

Theorem 12 (Transitivity Theorem)**.** *Let $Q \subseteq D^n$ be a relation that is hyperelementary over an interpreted language $\mathcal{L}_\mathcal{M}$ with domain $\mathcal{D}$. Let $\mathcal{L}^+_\mathcal{M}$ be the result of adding a new relation symbol χ^n to $\mathcal{L}_\mathcal{M}$, interpreted as having the extension Q, and let $R \subseteq \mathcal{D}^m$ be a relation that is inductive over $\mathcal{L}^+_\mathcal{M}$. Then R is inductive over $\mathcal{L}_\mathcal{M}$.*

In short: Expanding an interpreted language by adding a symbol for a hyperelementary relation does not allow one to define any new inductive relations.

4. Truth and Inductive Definitions: Tarski

In this section, we shall show how the consistency of Tarski's theory of truth for a given language $\mathcal{L}_\mathcal{M}$ can be proven using the theory of inductive definitions.

As above, let $\mathcal{L}_\mathcal{M}$ be an interpreted language consisting of a language $\mathcal{L}$ and an interpretation $\mathcal{M}$ of $\mathcal{L}$ with domain $\mathcal{D}$. Let $\mathcal{T}_{\mathcal{L}_\mathcal{M}}$ be its 'diagram', that is, the set of all true sentences of $\mathcal{L}_\mathcal{M}$. We assume that $\mathcal{T}_{\mathcal{L}_\mathcal{M}}$ interprets basic syntax: In particular, we can code finite sequences of elements of $\mathcal{D}$ by means of elements of $\mathcal{D}$; more precisely, we are assuming that notions such as atomic formula of $\mathcal{L}$, formula of $\mathcal{L}$, and the like, are $\mathcal{L}_\mathcal{M}$-elementary, and that $\mathcal{T}_{\mathcal{L}_\mathcal{M}}$ proves their basic properties. We use 'Seq(a, σ)' to mean: σ is a sequence that assigns values to all variables free in the formula (with Gödel number) a; '$\mathcal{L}$-AtForm(a)', that a is an atomic formula of $\mathcal{L}$; $\mathcal{L}$-Form(a), that a is a formula of $\mathcal{L}$.

We will also assume, mostly for convenience, that satisfaction for atomic formulae, SatAt$_\mathcal{M}(a, \sigma)$, is $\mathcal{L}_\mathcal{M}$-elementary, that is, that satisfaction for atomic formulae is definable by means of a formula of $\mathcal{L}_\mathcal{M}$.[9]

[9] In light of the transitivity theorem, it would actually be sufficient for our purposes to assume that SatAt$_\mathcal{M}(a, \sigma)$ is hyperelementary.

Let me insert a note on the meanings of expressions such as '$a = \underline{\neg}b$', used below. Of course, our formal language does not actually contain such expressions. By assumption, though, it does contain a formula $\text{neg}(a, b)$ meaning: a is (the code of a formula that is) a negation of (the formula with code) b. That is also what '$a = \underline{\neg}b$' means, though in using this notation we are assuming that only formulae have negations; that every formula has a negation; and that no formula has more than one negation. We are thus assuming that $\mathcal{T}_{\mathcal{L}_\mathcal{M}}$ proves such basic syntactic facts.

Theorem 13 (Tarski)**.** *Satisfaction for $\mathcal{L}_\mathcal{M}$ is not elementary, but it is hyperelementary.*

Note what this means: That satisfaction is not elementary means that it is not defined by any formula of $\mathcal{L}_\mathcal{M}$. This follows from Tarski's theorem, which will not be proven here.[10] That it is hyperelementary means that it is both inductive and co-inductive, which of course implies that both the set of truths and the set of falsehoods are inductive.

Note: *We are assuming we understand what satisfaction is*. Take it to be defined as Tarski defined it. Our goal is to prove facts *about* satisfaction, in particular, to prove that it can be defined by means of an inductive definition of the sort discussed in the previous section.

Proof of Theorem 13. Select arbitrary terms—which we shall abbreviate '$\bot$' and '$\top$'—of $\mathcal{L}_\mathcal{M}$ such that $\ulcorner \bot \neq \top \urcorner$ is true. Define $\text{Val}_\mathcal{M}(a, \sigma, t)$ as:

$$\mathcal{L}\text{-Form}(a) \wedge \text{Seq}(a, \sigma)) \wedge \{[t = \top \wedge \text{Sat}_\mathcal{M}(a, \sigma)] \vee [t = \bot \wedge \text{Sat}_\mathcal{M}(\underline{\neg}a, \sigma)]\}.$$

(Thus, $\top$ represents true and $\bot$ represents false.) We shall prove that $\{< a, \sigma, t >: \text{Val}_\mathcal{M}(a, \sigma, t)\}$ is a fixed-point over $\mathcal{L}_\mathcal{M}$. Since $\text{Sat}_\mathcal{M}(a, \sigma) \equiv \text{Val}_\mathcal{M}(a, \sigma, \top)$, it follows that $\{< a, \sigma >: \text{SatAt}_\mathcal{M}(a, \sigma)\}$ is inductive. Since

$$\begin{aligned}\text{Sat}_\mathcal{M}(a, \sigma) &\equiv \mathcal{L}\text{-Form}(a) \wedge \neg\text{Val}_\mathcal{M}(a, \sigma, \bot) \\ &\equiv \neg[\neg\mathcal{L}\text{-Form}(a) \vee \text{Val}_\mathcal{M}(a, \sigma, \bot)]\end{aligned}$$

it will be co-inductive and so hyperelementary.

Define $\phi(a, \sigma, t, \chi^3)$ as follows:

$$\begin{aligned}&\mathcal{L}\text{-Form}(a) \wedge \text{Seq}(a, \sigma) \wedge \{ \\ &\qquad [\mathcal{L}\text{-AtForm}(a) \wedge \text{SatAt}_\mathcal{M}(a, \sigma) \wedge t = \top] \vee \\ &\qquad [\mathcal{L}\text{-AtForm}(a) \wedge \neg\text{SatAt}_\mathcal{M}(a, \sigma) \wedge t = \bot] \vee \\ &\qquad \exists b[a = \underline{\neg}b \wedge \chi(b, \sigma, \top) \wedge t = \bot] \vee \\ &\qquad \exists b[a = \underline{\neg}b \wedge \chi(b, \sigma, \bot) \wedge t = \top] \vee \\ &\qquad \exists b \exists c[a = b\underline{\vee}c \wedge (\chi(b, \sigma, \top) \vee \chi(c, \sigma, \top)) \wedge t = \top] \vee \\ &\qquad \exists b \exists c[a = b\underline{\vee}c \wedge (\chi(b, \sigma, \bot) \wedge \chi(c, \sigma, \bot)) \wedge t = \bot] \vee \\ &\qquad \exists b \exists i[a = \underline{\exists v_i}b \wedge \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \chi(b, \tau, \top)) \wedge t = \top] \vee \\ &\qquad \exists b \exists i[a = \underline{\exists v_i}b \wedge \forall \tau(\forall j \neq i(\sigma_j = \tau_j) \rightarrow \chi(b, \tau, \bot)) \wedge t = \bot]\}\end{aligned}$$

By inspection, ϕ is χ-positive. So it has a fixed point I_χ.

Note. Think of the three-place predicate $\chi(a, \sigma, t)$ as meaning: t is the truth-value of a under the assignment σ of values to variables. One might have thought we should simply use a two-place formula $\upsilon(a, \sigma)$, meaning, intuitively: a is true

[10] See my "Formal Background for Theories of Truth" for a simple exposition.

under the assignment σ of values to variables (or just: σ satisfies a). We would then instead define $\phi(a, \sigma, \upsilon^2)$ as:

$$\begin{aligned}\mathcal{L}\text{-Form}(a) \wedge Seq(a, \sigma) \wedge \{ & \\ & [\mathcal{L}\text{-AtForm}(a) \wedge \text{SatAt}_{\mathcal{M}}(a, \sigma)] \vee \\ & [\mathcal{L}\text{-AtForm}(a) \wedge \neg\text{SatAt}_{\mathcal{M}}(a, \sigma)] \vee \\ & \exists b[a = \underline{\neg} b \wedge \neg\upsilon(b, \sigma)] \vee \\ & \exists b \exists c[a = b \underline{\vee} c \wedge (\upsilon(b, \sigma) \vee \upsilon(c, \sigma))] \vee \\ & \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \upsilon(b, \tau))]\}\end{aligned}$$

But then ϕ would not be υ-positive—look at the clause for negation—and the theory of inductive definitions would not apply to it. In that respect, it is essential here to make use of truth-*values* and not just of a truth-*predicate*.

End of note.

We intend to show that $\text{Val}_{\mathcal{M}}(a, \sigma, t) \equiv \text{<}a, \sigma, t\text{>} \in I_\chi$. To do this, we show by induction on α that

(i) $\qquad \text{<}a, \sigma, t\text{>} \in I^\alpha_\chi \rightarrow \text{Val}_{\mathcal{M}}(a, \sigma, t)$

Since $I_\chi = \bigcup_\alpha I^\alpha_\chi$, then if $\text{<}a, \sigma, t\text{>} \in I_\chi$, also $\text{<}a, \sigma, t\text{>} \in I^\alpha_\chi$, for some α. So we may conclude that $\text{<}a, \sigma, t\text{>} \in I_\chi \rightarrow \text{Val}_{\mathcal{M}}(a, \sigma, t)$. We will then show that

(ii) $\qquad \text{Val}_{\mathcal{M}}(a, \sigma, t) \rightarrow \exists\alpha(\text{<}a, \sigma, t\text{>} \in I^\alpha_\chi)$

and conclude that $\text{Val}_{\mathcal{M}}(a, \sigma, t) \rightarrow \text{<}a, \sigma, t\text{>} \in I_\chi$.

We first prove (i). Obviously, this holds vacuously for $\alpha = 0$, since $I^0_\chi = \emptyset$. So suppose it holds for $\xi < \alpha$; we want to show it also holds for α. Now, $I^\alpha_\chi = \Sigma^{\mathcal{M}}_\chi(\bigcup_{\xi<\alpha} I^\xi_\chi)$, which—abusing notation slightly—is the set of $\text{<}a, \sigma, t\text{>}$ such that $\phi(a, \sigma, t, \chi^3)$ when χ^3 has $\bigcup_{\xi<\alpha} I^\xi_\chi$ for its extension. So, if $\text{<}a, \sigma, t\text{>} \in I^\alpha_\chi$, then $\mathcal{L}$-Form(a) and Seq(a, σ) and one of the disjuncts of ϕ must hold. Say it's the fourth, for illustration. Then $a = \underline{\neg} b$ and $\chi(b, \sigma, \top)$ and $t = \bot$, and we must show that $\text{Val}_{\mathcal{M}}(a, \sigma, \bot)$. Since $\chi(b, \sigma, \top)$, we have $\text{<}b, \sigma, \top\text{>} \in \bigcup_{\xi<\alpha} I^\xi_\chi$ and so, for some $\xi < \alpha$, $\text{<}b, \sigma, \top\text{>} \in I^\xi_\alpha$. So, by the induction hypothesis, $\text{Val}_{\mathcal{M}}(a, \sigma, \top)$, that is, $\text{Sat}_{\mathcal{M}}(b, \sigma)$. So, by the definition of satisfaction, $\neg\text{Sat}_{\mathcal{M}}(a, \sigma)$, whence $\text{Val}_{\mathcal{M}}(a, \sigma, \bot)$. The other cases are similar.

To prove (ii), we argue by induction on the complexity of the formula a that:

(*) $\qquad \forall\sigma\forall t[\text{Val}_{\mathcal{M}}(a, \sigma, t) \rightarrow \exists\alpha(\text{<}a, \sigma, t\text{>} \in I^\alpha_\chi)]$

Certainly this holds for atomic formulae. If $t = \top$, then $\text{Sat}_{\mathcal{M}}(a, \sigma)$ because $\text{SatAt}_{\mathcal{M}}(a, \sigma)$, in which case $\text{<}a, \sigma, \top\text{>} \in I^1_\chi$; if $t = \bot$, then $\neg\text{Sat}_{\mathcal{M}}(a, \sigma)$ because $\neg\text{SatAt}_{\mathcal{M}}(a, \sigma)$, in which case $\text{<}a, \sigma, \bot\text{>} \in I^1_\chi$.

So suppose (*) holds for formulae of lesser complexity than a. By the induction hypothesis, for each such formula b and for each t and τ, if $\text{Val}_{\mathcal{M}}(b, \tau, t)$, then there is a ξ such that $\text{<}b, \tau, t\text{>} \in I^\xi_\chi$. Let α be the supremum of all such ξ.[11] Then since X is monotone, $\text{<}b, \tau, t\text{>} \in I^\alpha_\chi$. That is, for all b of complexity less than that of a and for all t and τ, if $\text{Val}_{\mathcal{M}}(b, \tau, t)$, then $\text{<}b, \tau, t\text{>} \in I^\alpha_\chi$.

So we now need to check the various possibilities for the kind of formula a might be and for the value t might take. We'll do two cases. First, suppose $a = \underline{\neg} b$ and $\text{Val}_{\mathcal{M}}(a, \sigma, t)$. If $t = \top$, then $\text{Sat}_{\mathcal{M}}(a, \sigma)$; so $\neg\text{Sat}_{\mathcal{M}}(b, \sigma)$, and so $\text{Val}_{\mathcal{M}}(b, \sigma, \bot)$.

[11] The supremum—that is, the least α greater than or equal to all such ξ—exists because the $\text{<}b, \tau, t\text{>}$ form a set.

By the results of the previous paragraph, $<b, \sigma, \bot> \in I_\chi^\alpha$. But then $\phi(a, \sigma, \top, \chi^3)$ will hold when χ^3 takes I_χ^α for its extension, since the third disjunct in the definition of ϕ will hold. So $<a, \sigma, \top> \in I_\chi^{\alpha+1}$. And if $t = \bot$, just swap $\bot$ and $\top$ in the foregoing.

Suppose that $a = \underline{\exists v_1} b$ and that $\text{Val}_\mathcal{M}(a, \sigma, t)$. Let $t = \top$. So $\text{Sat}_\mathcal{M}(a, \sigma)$ and hence, for some τ such that $\forall j \neq i(\sigma_j = \tau_j)$, $\text{Sat}_\mathcal{M}(b, \sigma)$ and so $\text{Val}_\mathcal{M}(b, \tau, \top)$. So $<b, \tau, \top> \in I_\chi^\alpha$. But then we have $\phi(a, \sigma, \top, \chi^3)$ when χ^3 takes I_χ^α for its extension, since the seventh of the disjuncts above will hold. It follows that $<a, \sigma, \top> \in I_\chi^{\alpha+1}$. If $t = \bot$, then $\neg\text{Sat}_\mathcal{M}(a, \sigma)$, so for all τ such that $\forall j \neq i(\sigma_j = \tau_j)$, $\neg\text{Sat}_\mathcal{M}(b, \tau)$ and so $\text{Val}_\mathcal{M}(b, \tau, \bot)$. But then $<b, \tau, \bot> \in I_\chi^\alpha$ and so $\phi(a, \sigma, \bot, \chi^3)$ when χ^3 takes I_χ^α for its extension, since the eighth disjunct holds, so $<a, \sigma, \bot> \in I_\chi^{\alpha+1}$. □

Corollary 14. *For every formula a and sequence σ, either $<a, \sigma, 1> \in I_\chi^\mathcal{M}$ or $<a, \sigma, \bot> \in I_\chi^\mathcal{M}$, but not both.*

Proof. This follows immediately from the preceding, since either $\text{Sat}_\mathcal{M}(a, \sigma)$ or $\neg\text{Sat}_\mathcal{M}(a, \sigma)$. But it can also be proven by induction on the complexity of a. □

Theorem 15 (Tarski)**.** *Let $\mathcal{T}$ be any theory in $\mathcal{L}$ meeting these two assumptions:*

(1) *$\mathcal{T}$ interprets basic syntax;*

(2) *There is a formula $\text{SatAt}_\mathcal{M}(a, \sigma)$ of $\mathcal{L}$ such that, for every atomic formula ϕ, $\mathcal{T}$ proves $\text{SatAt}_\mathcal{M}(\ulcorner\phi(x_1, \ldots, x_n)\urcorner, \sigma) \equiv \phi(\sigma_1, \ldots, \sigma_n)$.*

Let $\mathcal{T}_{sat}$ be the theory in $\mathcal{L} +$ 'sat' that is the result of adding to $\mathcal{T}$ the single axiom:

$$\begin{aligned} sat(a, \sigma) \equiv\ & \mathcal{L}\text{-Form}(a) \wedge \text{Seq}(a, \sigma) \wedge \{ \\ & [\mathcal{L}\text{-AtForm}(a) \wedge \text{SatAt}_\mathcal{M}(a, \sigma) \vee \\ & \exists b[a = \underline{\neg} b \wedge \neg sat(b, \sigma) \vee \\ & \exists b \exists c[a = b \underline{\vee} c \wedge (sat(b, \sigma) \vee sat(c, \sigma)] \vee \\ & \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists t \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge sat(b, \tau))]\}, \end{aligned}$$

Then $\mathcal{T}_{sat}$ is a conservative extension of $\mathcal{T}$ and so is consistent if $\mathcal{T}$ is.

Proof. Let $\mathcal{M}$ be any model of $\mathcal{T}$.

We show first that condition (2) implies that $\text{SatAt}_\mathcal{M}(a, \sigma)$ defines satisfaction in $\mathcal{M}$ for atomic formulae of $\mathcal{L}$. Let $\phi(x_1, \ldots, x_n)$ be an atomic formula of $\mathcal{L}$. Since $\mathcal{M}$ is a model of $\mathcal{T}$, $\mathcal{M} \models \forall\sigma[\text{SatAt}_\mathcal{M}(\ulcorner\phi(x_1, \ldots, x_n)\urcorner, \sigma) \equiv \phi(\sigma_1, \ldots, \sigma_n)]$. Now let s be a finite sequence of elements of the domain of $\mathcal{M}$ and let σ be a closed term coding s. Either $\mathcal{M}, s \models \phi(x_1, \ldots, x_n)$ or not. If so, then $\mathcal{M} \models \text{SatAt}_\mathcal{M}(\ulcorner\phi(x_1, \ldots, x_n)\urcorner, \sigma)$; if not, then $\mathcal{M} \models \neg\text{SatAt}_\mathcal{M}(\ulcorner\phi(x_1, \ldots, x_n)\urcorner, \sigma)$. Hence, $\mathcal{M}, s \models \phi(x_1, \ldots, x_n)$ iff $\mathcal{M} \models \text{SatAt}_\mathcal{M}(\ulcorner\phi(x_1, \ldots, x_n)\urcorner, \sigma)$, and $\text{SatAt}_\mathcal{M}(a, \sigma)$ defines $\mathcal{M}$-satisfaction for atomic formulae of $\mathcal{L}$.

Let $I_{\mathcal{M},\phi}$ be the fixed point over $\mathcal{M}$ of the formula $\phi(a, \sigma, t, \chi^3)$ defined in the proof of Theorem 13. Expand $\mathcal{M}$ to a model $\mathcal{M}_\chi$ of $\mathcal{L} + \chi$ by assigning $I_{\mathcal{M},\phi}$ as the extension of 'χ'; rewrite 'χ' as 'Val' for clarity.

Now consider the result of adding to $\mathcal{T}$ the following three axioms:

$$\mathcal{L}\text{-Form}(a) \wedge \text{Seq}(a, \sigma) \rightarrow \text{Val}_\mathcal{M}(a, \sigma, \top) \vee \text{Val}_\mathcal{M}(a, \sigma, 0)$$

$$\neg[\text{Val}_\mathcal{M}(a, \sigma, \top) \wedge \text{Val}_\mathcal{M}(a, \sigma, 0)]$$

$$\begin{aligned}\mathbf{Val}_{\mathcal{M}}(a,\sigma,t) \equiv\ & \mathcal{L}\text{-}\mathbf{Form}(a) \wedge \mathbf{Seq}(a,\sigma) \wedge \{\\ & [\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \mathbf{SatAt}_{\mathcal{M}}(a,\sigma) \wedge t = 1\top] \vee\\ & [\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \neg\mathbf{SatAt}_{\mathcal{M}}(a,\sigma) \wedge t = \bot] \vee\\ & \exists b[a = \underline{\neg} b \wedge \mathbf{Val}_{\mathcal{M}}(b,\sigma,1) \wedge t = \bot] \vee\\ & \exists b[a = \underline{\neg} b \wedge \mathbf{Val}_{\mathcal{M}}(b,\sigma,0) \wedge t = \top] \vee\\ & \exists b \exists c[a = b \underline{\vee} c \wedge (\mathbf{Val}_{\mathcal{M}}(b,\sigma,\top) \vee \mathbf{Val}_{\mathcal{M}}(c,\sigma,\top)) \wedge t = \top] \vee\\ & \exists b \exists c[a = b \underline{\vee} c \wedge (\mathbf{Val}_{\mathcal{M}}(b,\sigma,\bot) \wedge \mathbf{Val}_{\mathcal{M}}(c,\sigma,\bot)) \wedge t = \bot] \vee\\ & \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \mathbf{Val}_{\mathcal{M}}(b,\tau,\top)) \wedge t = \top] \vee\\ & \exists b \exists i[a = \underline{\exists v_i} b \wedge \forall \tau(\forall j \neq i(\sigma_j = \tau_j) \rightarrow \mathbf{Val}_{\mathcal{M}}(b,\tau,\bot)) \wedge t = \bot]\}\end{aligned}$$

Call this theory $\mathcal{T}_{Val}$. It is a conservative extension of $\mathcal{T}$, by theorem 5 and corollary 14.

Simple logical manipulations applied to the third axiom give us:

$$\begin{aligned}\mathbf{Val}_{\mathcal{M}}(a,\sigma,1) \equiv\ & \mathcal{L}\text{-}\mathbf{Form}(a) \wedge \mathbf{Seq}(a,\sigma) \wedge \{\\ & [\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \mathbf{SatAt}_{\mathcal{M}}(a,\sigma) \wedge t = \top] \vee\\ & \exists b[a = \underline{\neg} b \wedge \mathbf{Val}_{\mathcal{M}}(b,\sigma,\bot) \wedge t = \top] \vee\\ & \exists b \exists c[a = b \underline{\vee} c \wedge (\mathbf{Val}_{\mathcal{M}}(b,\sigma,\top) \vee \mathbf{Val}_{\mathcal{M}}(c,\sigma,\top)) \wedge t = \top] \vee\\ & \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \mathbf{Val}_{\mathcal{M}}(b,\tau,\top)) \wedge t = \top]\}\end{aligned}$$

Now define $sat(a,\sigma)$ as: $\mathbf{Val}_{\mathcal{M}}(a,\sigma,1)$. The first two axioms then yield:

$$\mathcal{L}\text{-}\mathbf{Form}(x) \wedge \mathbf{Seq}(a,\sigma) \rightarrow [\neg sat(x,\sigma) \equiv \neg\mathbf{Val}_{\mathcal{M}}(x,\sigma,0)]$$

Standard logical manipulations then show that the axiom mentioned in the theorem follows logically from the preceding and the definition of 'sat'. So $\mathcal{T}_{sat}$ is definitionally equivalent to $\mathcal{T}_{Val}$, which is itself a conservative extension of $\mathcal{T}$. So $\mathcal{T}_{sat}$ is a conservative extension of $\mathcal{T}$. □

Corollary 16. *Let $\mathcal{T}$ be any theory meeting the assumptions stated in Theorem 15. Then the theory $\mathcal{T}_{Tarski}$ in $\mathcal{L}$ + 'sat', the result of adding to $\mathcal{T}$ the axioms:*

(1) $sat(a,\sigma) \rightarrow \mathcal{L}\text{-}\mathbf{Form}(a) \wedge \mathbf{Seq}(a,\sigma)$
(2) $\mathcal{L}\text{-}\mathbf{AtForm}(a) \rightarrow [sat(a,\sigma) \equiv \mathbf{SatAt}_{\mathcal{M}}(a,\sigma)]$
(3) $sat(\ulcorner \neg A \urcorner, \sigma) \equiv \neg sat(A,\sigma)$
(4) $sat(\ulcorner A \vee B \urcorner, \sigma) \equiv sat(A,\sigma) \vee sat(B,\sigma)$
(5) $sat(\ulcorner \exists v_i A \urcorner, \sigma) \equiv \exists \tau[\forall j \neq i(\sigma_j = \tau_j) \wedge sat(A,\tau)]$

is a conservative extension of $\mathcal{T}$ and so is consistent if $\mathcal{T}$ is.

Proof. Each of (1)–(5) is provable in $\mathcal{T}_{sat}$, so $\mathcal{T}_{Tarski}$ is a sub-theory of $\mathcal{T}_{sat}$. □

Note that the fact that $\mathcal{T}_{sat}$ and $\mathcal{T}_{Tarski}$ are conservative extensions of $\mathcal{T}$ depends essentially upon the fact that 'sat' is not in the language of $\mathcal{T}$ and so does not occur in any axiom of the theory $\mathcal{T}$. In particular, it may not appear in formulae instantiating $\mathcal{T}$'s axiom schemes, should it have any.[12] If we do allow 'sat' to be used in that way, then $\mathcal{T}_{sat}$ and $\mathcal{T}_{Tarski}$ are not conservative extensions of $\mathcal{T}$ but will often prove the consistency of $\mathcal{T}$.

[12] That is, we are not thinking of $\mathcal{T}$ as a *schematic theory* in Feferman's sense.

5. TRUTH AND INDUCTIVE DEFINITIONS: KRIPKE

In this section, we shall show how the consistency of Kripke's theory of truth, for an interpreted language $\mathcal{L}_\mathcal{M}$, can be proven using the theory of inductive definitions. In fact, the changes needed to get Kripke's theory rather than Tarski's are surprisingly minimal.

Let $\mathcal{L}_\mathcal{M}$ be an interpreted language sufficient for syntax; let $\mathcal{L}^T_\mathcal{M}$ be the result of adding a single one-place predicate T intended to mean: is true, to $\mathcal{L}_\mathcal{M}$. We want to produce a truth-definition for $\mathcal{L}^T_\mathcal{M}$. We know, basically, what to do with the vocabulary of $\mathcal{L}_\mathcal{M}$. But what should the clause for the predicate T be like? Well, we want '$T(\mathbf{n})$' to be true just in case $\mathbf{n}$ is the Gödel number of a true sentence. We can think of this as part of an inductive specification of the extension of the predicate T: Once a sentence $\mathbf{n}$ has been deemed true, at the next stage of the induction, the sentence '$T(\mathbf{n})$' will be deemed true. On the other hand, we do *not* want '$T(\mathbf{n})$' to be deemed *false* simply because the sentence $\mathbf{n}$ has not yet been deemed true. We want '$T(\mathbf{n})$' to be deemed false only when the sentence $\mathbf{n}$ has itself been deemed false.

We can incorporate this idea into the inductive definition given earlier, though we shall do so for the general case of satisfaction, not just for truth. As before, let $\text{SatAt}_\mathcal{M}(a, \sigma)$ be a formula of $\mathcal{L}$ that defines satisfaction in $\mathcal{M}$ for atomic formulae of $\mathcal{L}$; $\mathcal{L}$-$\text{AtForm}(a)$, the notion of an atomic formula of $\mathcal{L}$; and $\mathcal{L}$-$\text{Form}(a)$, that of a formula. We assume further that there are formulae $\text{Term}(t)$ and $\text{den}_\mathcal{M}(t, \sigma, n)$ in $\mathcal{L}$ that define: a is a formula of $\mathcal{L}$; t is a term of $\mathcal{L}$, and: t denotes n under σ, respectively.[13] Further still, we assume that the theories we shall be considering below prove the basic facts about all of these notions. In particular, we assume that in all such theories we can prove:

$$\text{SatAt}_\mathcal{M}(\ulcorner \phi(x_1, \ldots, x_n) \urcorner, \sigma) \equiv \phi(\sigma_1, \ldots, \sigma_n)$$

for atomic formula ϕ and:

$$\text{den}_\mathcal{M}(\ulcorner t(x_1, \ldots, x_n) \urcorner, \sigma, n) \equiv t(\sigma_1, \ldots, \sigma_n) = n$$

for each term t. Such a theory might be called 'adequate'.

For convenience, we shall let $\text{den}_\mathcal{M}(t, \sigma)$ abbreviate: $\iota n(\text{den}_\mathcal{M}(t, \sigma, n))$, the description being proper since every term provably has a unique denotation (in adequate theories).[14]

As before, we now expand the language $\mathcal{L}_\mathcal{M}$ by adding a three-place predicate χ. Now let $\phi(a, \sigma, t, \chi^3)$ be the formula in Figure 5.1 on page 14. Note that this is the same definition of ϕ as used above, except for the last two disjuncts, which govern sentences of the form: $\chi(a, \sigma, t)$. Here, $a = \ulcorner \chi(t, u, 1) \urcorner$ means: a is (the code of) the formula: $\chi(\mathbf{t}, \mathbf{u}, 1)$, where $\mathbf{t}$ and $\mathbf{u}$ are the numerals denoting the numbers t and u, respectively.

What does this definition have to do with Kripke's? Kripke takes the truth-predicate to have both an *extension* and an *anti-extension*. As in the above discussion of Tarski, we are using a three-place predicate $\chi(a, \sigma, t)$, which we can think of as meaning: t is the truth-value of a under the assignment σ of values to variables. We may thus regard $\chi(a, \sigma, \top)$ as meaning: <a, σ> is in the extension of χ; and $\chi(a, \sigma, \bot)$ as meaning: <a, σ> is in the anti-extension of χ. The fact that the logic is three-valued is reflected in the fact that not every formula will

[13] One could take $\text{den}_\mathcal{M}(t, \sigma, n)$ to abbreviate: $\text{SatAt}_\mathcal{M}(\ulcorner t = n \urcorner, \sigma)$.

[14] That is, each formula $\ldots \text{den}_\mathcal{M}(t, \sigma) \ldots$ abbreviates the corresponding formula: $\exists n[\text{den}_\mathcal{M}(t, \sigma, n) \wedge \ldots n \ldots]$.

$$
\begin{aligned}
&\mathcal{L}\text{-}\mathbf{Form}(a) \wedge \mathbf{Seq}(a,\sigma) \wedge \{\\
&\qquad [\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \mathbf{SatAt}_{\mathcal{M}}(a,\sigma) \wedge t = \top] \vee\\
&\qquad [\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \neg\mathbf{SatAt}_{\mathcal{M}}(a,\sigma) \wedge t = \bot] \vee\\
&\qquad \exists b[a = \underline{\neg} b \wedge \chi(b,\sigma,\top) \wedge t = \bot] \vee\\
&\qquad \exists b[a = \underline{\neg} b \wedge \chi(b,\sigma,0) \wedge t = \top] \vee\\
&\qquad \exists b \exists c[a = b \underline{\vee} c \wedge (\chi(b,\sigma,\top) \vee \chi(c,\sigma,\top)) \wedge t = \top] \vee\\
&\qquad \exists b \exists c[a = b \underline{\vee} c \wedge (\chi(b,\sigma,0) \wedge \chi(c,\sigma,\bot)) \wedge t = \bot] \vee\\
&\qquad \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \chi(b,\tau,\top)) \wedge t = \top] \vee\\
&\qquad \exists b \exists i[a = \underline{\exists v_i} b \wedge \forall \tau(\forall j \neq i(\sigma_j = \tau_j) \rightarrow \chi(b,\tau,\bot)) \wedge t = \bot] \vee\\
&\qquad \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a = \ulcorner \chi(t,u,\top) \urcorner \wedge\\
&\qquad\qquad \chi(\mathbf{den}_{\mathcal{M}}(t,\sigma), \mathbf{den}_{\mathcal{M}}(u,\sigma), 1) \wedge t = \top] \vee\\
&\qquad \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a = \ulcorner \chi(t,u,\bot) \urcorner \wedge\\
&\qquad\qquad \chi(\mathbf{den}_{\mathcal{M}}(t,\sigma), \mathbf{den}_{\mathcal{M}}(u,\sigma), \bot) \wedge t = \bot]\}
\end{aligned}
$$

FIGURE 5.1. Kripke's Definition of Satisfaction

turn up in the 'extension' or 'anti-extension' in the minimal fixed point: We shall not, that is to say, be able to prove an analogue of Corollary 14. The fact that we are working with the *Strong* Kleene scheme is reflected in the character of the clauses for disjunction. To get the Weak Kleene scheme, define:

$$TV(x,\sigma) \equiv \exists y(\chi(x,\sigma,y)),$$

meaning: x has a truth-value under σ, and replace the first clause for disjunction with this one:

$$\exists b \exists c[a = b \underline{\vee} c \wedge TV(b,\sigma) \wedge TV(c,\sigma) \wedge (\chi(b,\sigma,1) \vee \chi(c,\sigma,1)) \wedge t = 1]$$

A similar change needs to be made to the clause for the existential quantifier.

One can make the definition look more like Kripke's by adding *two* two-place predicates τ and φ, intended to express truth and falsity or, more generally, satisfaction and non-satisfaction. Consider, then, the two formulae in Figure 5.2 on page 15, $TSat(a,\sigma,\tau^2,\varphi^2)$ and $FSat(a,\sigma,\tau^2,\varphi^2)$. Here τ represents satisfaction; φ, non-satisfaction. It follows from Lemma 11 that there are simultaneous fixed points I_{TSat} and I_{FSat} of $TSat$ and $FSat$: These will be the extension and anti-extension of the truth-predicate in Kripke's construction.

The above definition of the formula $\phi(a,\sigma,t,\chi^3)$ is essentially what one gets by applying the procedure described in the proof of Theorem 9 to $TSat(a,\sigma,\tau^2,\varphi^2)$ and $FSat(a,\sigma,\tau^2,\varphi^2)$. One could debate whether the pair of formulae or the single formula yields a more natural theory. Formally, however, it is a bit easier to work with the pair of definitions, so we shall use them. But it is important to appreciate the mathematical relationship between Kripke's treatment and Tarski's: Though their definitions of truth *look* very different, they are, both mathematically and conceptually, very similar.

There is another difference between this definition of satisfaction and Kripke's. In Kripke's construction, one begins by putting all true sentences of the base language $\mathcal{L}_{\mathcal{M}}$ into the extension. To get a definition more like Kripke's, one can replace the clauses of the $TSat$ and $FSat$ concerning atomic formulae of $\mathcal{L}$ with

$$
\begin{aligned}
&\mathcal{L}\text{-Form}(a)\wedge\text{Seq}(a,\sigma)\wedge\\
&\qquad\{[\mathcal{L}\text{-AtForm}(a)\wedge\text{SatAt}_{\mathcal{M}}(a,\sigma)]\vee\\
&\qquad\exists b[a=\underline{\neg}b\wedge\varphi(b,\sigma)]\vee\\
&\qquad\exists b\exists c[a=b\underline{\vee}c\wedge(\tau(b,\sigma)\vee\tau(c,\sigma))]\vee\\
&\qquad\exists b\exists i[a=\underline{\exists v_i}b\wedge\exists\upsilon(\forall j\neq i(\sigma_j=\upsilon_j)\wedge\tau(b,\upsilon))]\vee\\
&\qquad\exists t\exists u[\text{Term}(t)\wedge\text{Term}(u)\wedge a=\ulcorner\tau(t,u)\urcorner\wedge\\
&\qquad\qquad\tau(\text{den}_{\mathcal{M}}(t,\sigma),\text{den}_{\mathcal{M}}(u,\sigma))]\vee\\
&\qquad\exists t\exists u[\text{Term}(t)\wedge\text{Term}(u)\wedge a=\ulcorner\varphi(t,u)\urcorner\wedge\\
&\qquad\qquad\varphi(\text{den}_{\mathcal{M}}(t,\sigma),\text{den}_{\mathcal{M}}(u,\sigma))]\}
\end{aligned}
$$

$$
\begin{aligned}
&\mathcal{L}\text{-Form}(a)\wedge\text{Seq}(a,\sigma)\wedge\\
&\qquad\{[\mathcal{L}\text{-AtForm}(a)\wedge\neg\text{SatAt}_{\mathcal{M}}(a,\sigma)]\vee\\
&\qquad\exists b[a=\underline{\neg}b\wedge\tau(b,\sigma)]\vee\\
&\qquad\exists b\exists c[a=b\underline{\vee}c\wedge\varphi(b,\sigma)\wedge\varphi(c,\sigma)]\vee\\
&\qquad\exists b\exists i[a=\underline{\exists v_i}b\wedge\forall\upsilon(\forall j\neq i(\sigma_j=\upsilon_j)\rightarrow\varphi(b,\upsilon))]\vee\\
&\qquad\exists t\exists u[\text{Term}(t)\wedge\text{Term}(u)\wedge a=\ulcorner\varphi(t,u)\urcorner\wedge\\
&\qquad\qquad\tau(\text{den}_{\mathcal{M}}(t,\sigma),\text{den}_{\mathcal{M}}(u,\sigma))]\vee\\
&\qquad\exists t\exists u[\text{Term}(t)\wedge\text{Term}(u)\wedge a=\ulcorner\tau(t,u)\urcorner\wedge\\
&\qquad\qquad\varphi(\text{den}_{\mathcal{M}}(t,\sigma),\text{den}_{\mathcal{M}}(u,\sigma))]\}
\end{aligned}
$$

FIGURE 5.2. Kripke's Definition, Again

these clauses:

$$
\begin{aligned}
&\mathcal{L}\text{-Form}(a)\wedge sat_{\mathcal{L}_{\mathcal{M}}}(a,\sigma)\\
&\mathcal{L}\text{-Form}(a)\wedge\neg sat_{\mathcal{L}_{\mathcal{M}}}(a,\sigma)
\end{aligned}
$$

where $sat_{\mathcal{L}_{\mathcal{M}}}$ is a satisfaction predicate for $\mathcal{L}_{\mathcal{M}}$, defined as above. Since satisfaction for $\mathcal{L}_{\mathcal{M}}$ is hyperelementary over $\mathcal{L}_{\mathcal{M}}$, there is a sense in which this difference makes no difference: By Theorem 12, any relation inductive over $\mathcal{L}_{\mathcal{M}}+sat_{\mathcal{M}}$ is also inductive over $\mathcal{L}_{\mathcal{M}}$.

On the other hand, this difference does affect the 'levels' at which sentences get truth-values. On Kripke's treatment, the level of every (code of a) sentence of the base language is 1. This will not be so if we use the inductive definition given above: Only *atomic* sentences of $\mathcal{L}_{\mathcal{M}}$ will get truth-values at stage 1. Perhaps Kripke's way of doing things is preferable if one is interested only in the semantics of sentences containing the truth-predicate. But the definition above is more natural if we are not. This definition does not treat 'true' in any way peculiarly; 'true' is treated pretty much the same way any other atomic predicate is treated (a fact that is partly hidden by our use of 'SatAt$_{\mathcal{M}}(a,\sigma)$ rather than individual clauses for the various atomic expressions of $\mathcal{L}_{\mathcal{M}}$). That is to say: We are not thinking of *adding* a truth-predicate to a language for which truth has already been defined; we are defining satisfaction for a language that already contains a truth-predicate. We shall therefore stick with our present definition.

At the moment, then, what we have is an inductive definition of a pair of notions, satisfaction and non-satisfaction. We have thus now reached a point in our exposition comparable the one we had reached in our discussion of Tarksi after the proof of Theorem 13. We now want to consider what *formal theories* of truth we can prove to be consistent on the basis of these definitions. We begin by simply converting the formulae $TSat$ and $FSat$ into a pair of axioms.

Theorem 17. *Let $\mathcal{T}$ be a theory in a language $\mathcal{L}$. Let $\mathcal{T}_{Tsat,Fsat}$ be a theory in $\mathcal{L}$+$Tsat$+$Fsat$ containing the axioms of $\mathcal{T}$ plus the three additional axioms:*

$$\neg\exists x[Tsat(x,\sigma) \wedge Fsat(x,\sigma)]$$

$$\begin{aligned}
Tsat(a,\sigma) \equiv\ & \mathcal{L}\text{-}\mathbf{Form}(a) \wedge \mathbf{Seq}(a,\sigma) \wedge \\
& \{[\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \mathbf{SatAt}_{\mathcal{M}}(a,\sigma)] \vee \\
& \exists b[a = \underline{\neg} b \wedge Fsat(b,\sigma)] \vee \\
& \exists b \exists c[a = b \underline{\vee} c \wedge (Tsat(b,\sigma) \vee Tsat(c,\sigma))] \vee \\
& \exists b \exists i[a = \underline{\exists v_i} b \wedge \exists v(\forall j \neq i(\sigma_j = v_j) \wedge Tsat(b,v))] \vee \\
& \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a = \ulcorner Tsat(t,u) \urcorner \wedge \\
& \quad Tsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))] \vee \\
& \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a = \ulcorner Fsat(t,u) \urcorner \wedge \\
& \quad Fsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))]\}
\end{aligned}$$

$$\begin{aligned}
Fsat(a,\sigma) \equiv\ & Fmla(a) \wedge \mathbf{Seq}(a,\sigma) \wedge \\
& \{[\mathcal{L}\text{-}\mathbf{AtForm}(a) \wedge \neg\mathbf{SatAt}_{\mathcal{M}}(a,\sigma)] \vee \\
& \exists b[a = \underline{\neg} b \wedge Tsat(b,\sigma)] \vee \\
& \exists b \exists c[a = b \underline{\vee} c \wedge Fsat(b,\sigma) \wedge Fsat(c,\sigma)] \vee \\
& \exists b \exists i[a = \underline{\exists v_i} b \wedge \forall v(\forall j \neq i(\sigma_j = v_j) \rightarrow Fsat(b,v))] \vee \\
& \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a \wedge \ulcorner Fsat(t,u) \urcorner \wedge \\
& \quad Tsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))] \vee \\
& \exists t \exists u[\mathbf{Term}(t) \wedge \mathbf{Term}(u) \wedge a \wedge \ulcorner Tsat(t,u) \urcorner \wedge \\
& \quad Fsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))]\}
\end{aligned}$$

Then $\mathcal{T}_{Tsat,Fsat}$ is a conservative extension of $\mathcal{T}$ and so is consistent if $\mathcal{T}$ is. Moreover, the formulae:

$$Tsat(\underline{\neg} t,\sigma) \equiv Fsat(t,\sigma)$$
$$Fsat(\underline{\neg} t,\sigma) \equiv Tsat(t,\sigma)$$

are theorems of $\mathcal{T}_{Tsat,Fsat}$.

Proof. Let $\mathcal{M}$ be a model of $\mathcal{T}$. As mentioned, Lemma 11 guarantees that there are simultaneous fixed points $I^{\mathcal{M}}_{TSat}$ and $I^{\mathcal{M}}_{FSat}$ of the formulae $Tsat$ and $Fsat$ in Figure 5.2 on page 15. Expand $\mathcal{M}$ to an interpretation $\mathcal{M}^*$ of $\mathcal{L} + Tsat + Fsat$ by taking the extension of $Tsat$ to be $I^{\mathcal{M}}_{TSat}$ and that of $Fsat$ to be $I^{\mathcal{M}}_{FSat}$. Then $\mathcal{M}^*$ is a model of $\mathcal{T}_{Tsat,Fsat}$.

That $\neg\exists x[Tsat(x,\sigma) \wedge Fsat(x,\sigma)]$ holds in $\mathcal{M}^*$ is proven by induction on the stages of the inductive definition, that is, on α in I^{α}_{TSat} and I^{α}_{FSat}. The details are left as an exercise.

The final claim follows from the clauses for negation. □

Remark. We do *not* have $\neg Tsat(x,\sigma) \rightarrow Fsat(x,\sigma)$ or, equivalently, $Tsat(x,\sigma) \vee Fsat(x,\sigma)$.

We can simplify $\mathcal{T}_{Tsat,Fsat}$ substantially. We begin by stating an obvious corollary of the preceding theorem.

Corollary 18 (Kripke)**.** *Let $\mathcal{T}$ be a theory in a language $\mathcal{L}$. Let $\mathcal{T}_{TF}$ be the theory in $\mathcal{L}$+$Tsat$+$Fsat$ containing the axioms of $\mathcal{T}$ plus the universal closures of:*

(1) $Tsat(x,\sigma) \rightarrow \mathcal{L}\text{-Form}(x) \wedge \text{Seq}(a,\sigma)$
$Fsat(x,\sigma) \rightarrow \mathcal{L}\text{-Form}(x) \wedge \text{Seq}(a,\sigma)$

(2) $Tsat(\ulcorner \phi(x_1,\ldots,x_n) \urcorner, \sigma) \equiv \phi(\sigma_1,\ldots,\sigma_n)$ *and*
$Fsat(\ulcorner \phi(x_1,\ldots,x_n) \urcorner, \sigma) \equiv \neg\phi(\sigma_1,\ldots,\sigma_n)$*, for ϕ an atomic formula of $\mathcal{L}$*

(3) $Tsat(t \underline{\vee} u,\sigma) \equiv Tsat(t,\sigma) \vee Tsat(u,\sigma)$
$Fsat(t \underline{\vee} u,\sigma) \equiv Fsat(t,\sigma) \wedge Fsat(u,\sigma)$

(4) $Tsat(\underline{\exists v_i} t,\sigma) \equiv \exists\tau(\forall j \neq i(\sigma_j = \tau_j) \wedge Tsat(t,\tau))$
$Fsat(\underline{\exists v_i} t,\sigma) \equiv \forall\tau(\forall j \neq i(\sigma_j = \tau_j) \rightarrow Fsat(t,\tau))$

(5) $Tsat(\ulcorner Tsat(t,u) \urcorner,\sigma) \equiv Tsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))$
$Tsat(\ulcorner Fsat(t,u) \urcorner,\sigma) \equiv Fsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))$
$Fsat(\ulcorner Tsat(t,u) \urcorner,\sigma) \equiv Fsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))$
$Fsat(\ulcorner Fsat(t,u) \urcorner,\sigma) \equiv Tsat(\boldsymbol{den}_{\mathcal{M}}(t,\sigma), \boldsymbol{den}_{\mathcal{M}}(u,\sigma))$

(6) $\neg\exists x[Tsat(x,\sigma) \wedge Fsat(x,\sigma)]$

(7) $Tsat(\underline{\neg} t,\sigma) \equiv Fsat(t,\sigma)$
$Fsat(\underline{\neg} t,\sigma) \equiv Tsat(t,\sigma)$

$\mathcal{T}_{TF}$ has the same theorems as $\mathcal{T}_{Tsat,Fsat}$.

Note what the various clauses in (5) mean: The first means that a formula of the form $Tsat(t,u)$ is satisfied by a sequence σ if, and only if, the formula whose Gödel number is the denotation of t with respect to σ is satisfied by the sequence whose Gödel number is the denotation of u with respect to σ. Thus, $Tsat(t,u)$ 'says' that the formula (with Gödel number) t is 'truly' satisfied by the sequence (with Gödel number) u.

Proof. Axioms (1)–(5) of $\mathcal{T}_{TF}$ are easily provable in $\mathcal{T}_{Tsat,Fsat}$: They all follow immediately from appropriate clauses of its second and third axioms. Axioms (6) and (7) are already axioms of $\mathcal{T}_{Tsat,Fsat}$. So $\mathcal{T}_{TF}$ is a sub-theory of $\mathcal{T}_{Tsat,Fsat}$.

Conversely, the three axioms of $\mathcal{T}_{Tsat,Fsat}$ are provable in $\mathcal{T}_{TF}$. We prove just the second.

If we suppose that $Tsat(a,\sigma)$, then by axiom (1) of $\mathcal{T}_{TF}$, a is a formula and σ is a sequence. So a is either an atomic formula of $\mathcal{L}$, a negation, a disjunction, an existential quantification, or is of the form $Tsat(b,\tau)$ or $Fsat(b,\tau)$. The parts of axioms (2), (3), (4), (5), and (7) concerned with $Tsat$ then entail the corresponding disjuncts on the right-hand side of axiom (1). Conversely, if we suppose the right-hand side of the second axiom of $\mathcal{T}_{Tsat,Fsat}$, then, again, a is a formula and σ is a sequence; a must be of one of the specified forms; and the appropriate disjunct together with the corresponding axiom of $\mathcal{T}_{Tsat,Fsat}$ will entail $Tsat(a,\sigma)$. □

Even a casual examination of the axioms of $\mathcal{T}_{TF}$ reveals a massive and dissatisfying repetitiveness that is due to the fact that we are using distinct satisfaction and non-satisfaction axioms. But we can easily eliminate this redundancy: In virtue of (7), we can replace $Fsat(t,\sigma)$ throughout by $Tsat(\underline{\neg} t,\sigma)$. This gives us the following.

Theorem 19 (Kripke, Feferman)**.** *Let $\mathcal{T}$ be a theory in a language $\mathcal{L}$. Let $\mathcal{T}_{KF}$ be the theory in $\mathcal{L}$+Sat containing the axioms of $\mathcal{T}$ plus the universal closures of:*

(1) $Sat(x, \sigma) \rightarrow \mathcal{L}\text{-Form}(x) \wedge \text{Seq}(a, \sigma)$
(2) $Sat(\ulcorner \phi(x_1, \ldots, x_n) \urcorner, \sigma) \equiv \phi(\sigma_1, \ldots, \sigma_n)$*, for ϕ an atomic or negated atomic formula of $\mathcal{L}$*
(3) *a.* $Sat(t \underline{\vee} u, \sigma) \equiv Sat(t, \sigma) \vee Sat(u, \sigma)$
    b. $Sat(\underline{\neg}(t \underline{\vee} u), \sigma) \equiv Sat(\underline{\neg} t, \sigma) \wedge Sat(\underline{\neg} u, \sigma)$
(4) *a.* $Sat(\underline{\exists v_i} t, \sigma) \equiv \exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge Sat(t, \tau))$
    b. $Sat(\underline{\neg \exists v_i} t, \sigma) \equiv \forall \tau(\forall j \neq i(\sigma_j = \tau_j) \rightarrow Sat(\underline{\neg} t, \tau))$
(5) *a.* $Sat(\ulcorner \underline{Sat}(t, u) \urcorner, \sigma) \equiv Sat(\underline{\textit{den}}_{\mathcal{M}}(t, \sigma), \underline{\textit{den}}_{\mathcal{M}}(u, \sigma))$
    b. $Sat(\ulcorner \underline{\neg Sat}(t, u) \urcorner, \sigma) \equiv Sat(\underline{\neg \textit{den}}_{\mathcal{M}}(t, \sigma), \underline{\textit{den}}_{\mathcal{M}}(u, \sigma))$
(6) $\neg[Sat(x, \sigma) \wedge Sat(\underline{\neg} x, \sigma)]$
(7) $Sat(\underline{\neg \neg} t, \sigma) \equiv Sat(t, \sigma)$

$\mathcal{T}_{KF}$ is a definitional extension of $\mathcal{T}_{TF}$. It is therefore a conservative extension of $\mathcal{T}$ and so is consistent if $\mathcal{T}$ is.

The axioms just mentioned are called the *Kripke-Feferman satisfaction axioms*.

Proof. Define $Sat(a, \sigma)$ as: $Tsat(a, \sigma)$. Axioms (1)–(7) of $\mathcal{T}_{KF}$ then follow immediately from the corresponding axioms of $\mathcal{T}_{TF}$. □

The result just proven has a converse.

Theorem 20. *$\mathcal{T}_{TF}$ is a definitional extension of $\mathcal{T}_{KF}$.*

Proof. We need only define $Tsat(x, y)$ as: $Sat(x, y)$, and $Fsat(x, y)$ as: $Sat(\underline{\neg} x, y)$. □

The theory $\mathcal{T}_{KF}$ has a very nice property: It delivers half of Tarski's T-scheme.

Theorem 21. *For every formula ϕ of $\mathcal{L}$, $\mathcal{T}_{KF}$ proves:*

$$Sat(\ulcorner \phi(x_1, \ldots, x_n) \urcorner, \sigma) \rightarrow \phi(\sigma_1, \ldots, \sigma_n)$$

Proof. The proof is, of course, by induction on the complexity of formulas. In fact, the induction proceeds by showing that, for every formula ϕ, $\mathcal{T}_{KF}$ proves both of the followign:

$$Sat(\ulcorner \phi(x_1, \ldots, x_n) \urcorner, \sigma) \rightarrow \phi(\sigma_1, \ldots, \sigma_n)$$
$$Sat(\ulcorner \neg \phi(x_1, \ldots, x_n) \urcorner, \sigma) \rightarrow \neg \phi(\sigma_1, \ldots, \sigma_n)$$

We'll see the reason we need so to proceed shortly.

If ϕ is atomic, it is either an atomic formula of $\mathcal{L}$ or is of the form $Sat(t, u)$. If the former, the result follows from axiom (2) of $\mathcal{T}_{KF}$. So suppose ϕ is of the form $Sat(t, u)$. Then axiom (5a) gives us:

$$Sat(\ulcorner Sat(t, u) \urcorner, \sigma) \equiv Sat(\text{den}_{\mathcal{M}}(t, \sigma), \text{den}_{\mathcal{M}}(u, \sigma))$$

But we are assuming that $\mathcal{T}$ proves: $\text{den}_{\mathcal{M}}(t, \sigma) = t(\sigma_1, \ldots, \sigma_n)$. So

$$Sat(\text{den}_{\mathcal{M}}(t, \sigma), \text{den}_{\mathcal{M}}(u, \sigma)) \equiv Sat(t(\sigma_1, \ldots, \sigma_n), u(\sigma_1, \ldots, \sigma_n))$$

and similarly for the case of $\neg Sat(t, u)$.

If ϕ is a disjunction $\psi \vee \chi$, then we have:

$$Sat(\ulcorner \psi(x_1, \ldots, x_n) \vee \chi(x_1, \ldots, x_n) \urcorner, \sigma) \rightarrow$$
$$Sat(\ulcorner \psi(x_1, \ldots, x_n) \urcorner, \sigma) \vee Sat(\ulcorner \chi(x_1, \ldots, x_n) \urcorner, \sigma) \rightarrow$$
$$\psi(\sigma_1, \ldots, \sigma_n) \vee \chi(\sigma_1, \ldots, \sigma_n)$$

The first inference is by axiom (3a); the second, by the induction hypothesis. The case of negated disjunctions is similar.

So suppose ϕ is $\exists v_i \psi$; for convenience we assume $i = 1$. Then we have:

$$Sat(\ulcorner \exists v_1 \phi(v_1, \ldots, v_n) \urcorner, \sigma) \rightarrow$$
$$\exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge Sat(\ulcorner \phi(v_1, \ldots, v_n) \urcorner, \tau)) \rightarrow$$
$$\exists \tau(\forall j \neq i(\sigma_j = \tau_j) \wedge \phi(\tau_1, \ldots, \tau_n)) \rightarrow$$
$$\exists \tau \phi(\tau_1, \sigma_2, \ldots, \sigma_n) \rightarrow$$
$$\exists v_1 \phi(v_1, \sigma_2, \ldots, \sigma_n) \rightarrow$$

The first inference is by (4a); the second is by the induction hypothesis; the last two are simple logical manipulations. And again, the case of negated existentials is similar. □

Corollary 22. *In $\mathcal{T}_{KF}$, define a truth-predicate thus:*

$$Tr(t) \equiv Sent(t) \wedge Sat(t, \langle\rangle)$$

where $\langle\rangle$ is the empty sequence. Then for every sentence ϕ, $\mathcal{T}_{KF}$ proves: $Tr(\ulcorner \phi \urcorner) \rightarrow \phi$. Moreover, $\mathcal{T}_{KF}$ proves: $\neg \exists t[Tr(t) \wedge Tr(\underline{\neg} t)]$.

Proof. The first claim is immediate from the preceding corollary. The second follows from axiom (6) of $\mathcal{T}_{KF}$. □

Proposition 23. *$\mathcal{T}_{KF}$ is inconsistent with the Tr-introduction rule: $A \vdash Tr(\ulcorner A \urcorner)$.*

This rule is similar to the rule of necessitation in modal logic. What it says, in effect, is that all *theorems* of $\mathcal{T}_{KF}$ are true. Indeed, as we shall see, the only instance of this rule we need is one for a theorem of pure logic, not of $\mathcal{T}_{KF}$.

Proof. Let λ be a liar sentence. So $\mathcal{T}_{KF} \vdash \lambda \equiv \neg Tr(\ulcorner \lambda \urcorner)$. Then $\mathcal{T}_{KF}$ proves:

(1)	$Tr(\ulcorner \lambda \urcorner) \rightarrow \lambda$	by Corollary 22
(2)	$\lambda \rightarrow \neg Tr(\ulcorner \lambda \urcorner)$	since λ is a liar
(3)	$\neg Tr(\ulcorner \lambda \urcorner)$	from (1) and (2)
(4)	$Tr(\ulcorner \neg \lambda \urcorner) \rightarrow \neg \lambda$	by Corollary 22
(5)	$\neg \lambda \rightarrow Tr(\ulcorner \lambda \urcorner)$	since λ is a liar
(6)	$Tr(\ulcorner \neg \lambda \urcorner) \rightarrow Tr(\ulcorner \neg \lambda \urcorner) \wedge Tr(\ulcorner \lambda \urcorner)$	from (4) and (5)
(7)	$\neg[Tr(\ulcorner \neg \lambda \urcorner) \wedge Tr(\ulcorner \lambda \urcorner)]$	by Corollary 22
(8)	$\neg Tr(\ulcorner \neg \lambda \urcorner)$	from (6) and (7)
(9)	$Tr(\ulcorner \lambda \vee \neg \lambda \urcorner) \rightarrow Tr(\ulcorner \lambda \urcorner) \vee Tr(\ulcorner \neg \lambda \urcorner)$	by axiom (5) of KF
(10)	$\neg Tr(\ulcorner \lambda \vee \neg \lambda \urcorner)$	from (3), (8), and (9)

But $\mathcal{T}_{KF}$ proves: $\lambda \vee \neg \lambda$, trivially, as a truth of logic. So Tr-introduction would yield $Tr(\ulcorner \lambda \vee \neg \lambda \urcorner)$, and $\mathcal{T}_{KF}$ would be inconsistent. □

Corollary 24. *Let (E, A) be a fixed point of the Kripke construction over an interpreted language $\mathcal{L}_{\mathcal{M}}$ (i.e., what we get from the proof of Theorem 17). Then all axioms of $\mathcal{T}_{KF}$ are true when 'Sat' is interpreted by E. Conversely, if all axioms of $\mathcal{T}_{KF}$ are true when 'Sat' is interpreted by E, and A is defined as: $\langle t, \sigma \rangle \in A$ iff $\langle \underline{\neg} t, \sigma \rangle \in E$, then (E, A) is a fixed point of the Kripke construction over $\mathcal{L}_{\mathcal{M}}$.*

Proof. Left-to-right: Theorem 17 shows that, if (E, A) is a fixed point of the Kripke construction, then all axioms of $\mathcal{T}_{Tsat,Fsat}$ are true when '$Tsat$' is interpreted by E and '$Fsat$' is interpreted by A. But by Corollary 18 and Theorem 19, all axioms of $\mathcal{T}_{KF}$ are provable in $\mathcal{T}_{Tsat,Fsat}$ if '$Tsat$' is re-written as 'Sat', so all axioms of $\mathcal{T}_{KF}$ must be true when 'Sat' is interpreted by E.

Right-to-left: Suppose all axioms of $\mathcal{T}_{KF}$ are true when 'Sat' is interpreted by E and and define A as stated. By Theorem 20, if we re-write 'Sat' as '$Tsat$' and treat '$Fsat(a, \sigma)$' as an abbreviation of: $Tsat(\underline{\neg}a, \sigma)$, then all axioms of $\mathcal{T}_{TF}$ are theorems of $\mathcal{T}_{KF}$. So all axioms of $\mathcal{T}_{Tsat,Fsat}$ must be true when '$Tsat$' is interpreted by E and '$Fsat$' is treated as an abbreviation. □

6. Satisfaction-predicates and Three-valued Logic

The theories whose consistency we just proved were all *classical* theories: That is, the underlying logic of each of these theories was classical, two-valued logic. The theory thus does not yet look much like Kripke's theory of truth, which employs a three-valued logic, though there is an obvious sort of correspondence between the satisfaction and non-satisfaction predicates that appear above and the extension and anti-extension of the truth-predicate in Kripke's treatment, a correspondence we just exploited. We now show how to use it to recover Kripke's original theory.

Given an interpreted language $\mathcal{L}_{\mathcal{M}}$, define an interpreted language $\mathcal{L}^K_{\mathcal{M}}$ as follows. The atomic expressions of $\mathcal{L}^K_{\mathcal{M}}$ are to be those of $\mathcal{L}_{\mathcal{M}}$ plus a predicate $TSat_K(a, \sigma)$ intended to express satisfaction for $\mathcal{L}^K_{\mathcal{M}}$. Let the extension and anti-extension of $TSat_K(x, y)$ be the fixed points $I_{TSat,\mathcal{M}}$ and $I_{FSat,\mathcal{M}}$ and extend this to an interpretation of the whole of $\mathcal{L}^K_{\mathcal{M}}$ by means of the Strong Kleene scheme. That is, assume we have a definition of satisfaction for atomic formulae of $\mathcal{L}_{\mathcal{M}}$ and simultaneously define satisfaction and anti-satisfaction as follows:

σ satisfies ϕ if, and only if, either:

(1) ϕ is an atomic formula of $\mathcal{L}_{\mathcal{M}}$, and $\text{SatAt}_{\mathcal{M}}(\phi, \sigma)$;
(2) ϕ is a conjunction $\psi \wedge \chi$ and σ satisfies both ψ and χ;
(3) ϕ is a negation $\neg\psi$ and σ anti-satisfies ψ;
(4) ϕ is an existential quantification $\exists v_i \psi$ and there is some sequence τ agreeing with σ everywhere except, possibly, at i such that σ satisfies ψ;
(5) ϕ is $TSat_K(t, u)$, where t is the Gödel number of some formula ψ and u is the code of a sequence τ and τ satisfies ψ.

σ anti-satisfies ϕ if, and only if, either:

(1) ϕ is an atomic formula of $\mathcal{L}_{\mathcal{M}}$, and $\neg\text{SatAt}_{\mathcal{M}}(\phi, \sigma)$;
(2) ϕ is a conjunction $\psi \wedge \chi$ and σ anti-satisfies at least one of ψ and χ;
(3) ϕ is a negation $\neg\psi$ and σ satisfies ψ;
(4) ϕ is an existential quantification $\exists v_i \psi$ and for every sequence τ agreeing with σ everywhere except, possibly, at i, τ anti-satisfies ψ;
(5) ϕ is $TSat_K(t, u)$, where t is the Gödel number of some formula ψ and u is the code of a sequence τ and τ anti-satisfies ψ.

We say that ϕ is a true sentence of $\mathcal{L}^K_{\mathcal{M}}$ if it is satisfied by every sequence; a false sentence, if it is anti-satisfied by every sequence.

Theorem 25 (Kripke). *$TSat_K(a,\sigma)$ is a satisfaction predicate for $\mathcal{L}^K_{\mathcal{M}}$. That is: $TSat_K(a,\sigma)$ is a true sentence of $\mathcal{L}^K_{\mathcal{M}}$ just in case σ is (the code of) a sequence that satisfies the sentence (whose code is) a, and $TSat_K(a,\sigma)$ is a false sentence of $\mathcal{L}^K_{\mathcal{M}}$ just in case σ is (the code of) a sequence that anti-satisfies the sentence (whose code is) a.*[15]

Proof. The crucial observation is simply that the definition of satisfaction and anti-satisfaction just given transcribes the clauses of the simultaneous inductive definition of $I_{TSat,\mathcal{M}}$ and $I_{FSat,\mathcal{M}}$. The only difference is that we have omitted any reference to a formula $FSat_K(x,y)$, which might have been taken to express anti-satisfaction. The theorem can thus be proven by a straightforward inductive argument. □

To get an actual *formal theory* of truth for $\mathcal{L}^K_{\mathcal{M}}$, one must axiomatize the underlying three-valued logic of the theory. There are different ways to do this: See Kremer [3] and McGee [5] for a couple of different approaches. Once one has done that, however, the principles governing the truth- and satisfaction-predicates are easily stated. These will be the so-called T-rules:

$$\begin{aligned} A &\vdash T(\ulcorner A \urcorner) \\ \neg A &\vdash \neg T(\ulcorner A \urcorner) \\ T(\ulcorner A \urcorner) &\vdash A \\ \neg T(\ulcorner A \urcorner) &\vdash \neg A \end{aligned}$$

with corresponding rules for satisfaction.

7. Exercises

Exercise 1. Show that the converse of Lemma 8 also holds: If (σ) is true when σ^n is taken to have S as its extension, then S is a fixed point of $\Sigma^{\mathcal{M}}_{\phi}$.

Exercise 2. Complete the proof of Theorem 13 by doing the cases we omitted.

References

[1] Solomon Feferman. Toward useful type-free theories I. *Journal of Symbolic Logic*, 49:75–111, 1984.

[2] Alexander Kechris and Yiannis Moschovakis. Recursion in higher types. In John Barwise, editor, *A Handbook of Mathematical Logic*, pages 681–739. North-Holland Publishing, New York, 1977.

[3] Michael Kremer. Kripke and the logic of truth. *Journal of Philosophical Logic*, 17:225–78, 1988.

[4] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.

[5] Vann McGee. *Truth, Vagueness, and Paradox: An Essay on the Logic of Truth*. Hackett, 1990.

[6] Yiannis Moschovakis. *Elementary Induction on Abstract Structures*. North Holland Publishing, Amsterdam, 1974.

[15] One can state this more generally, to allow for the possibility that variables occur in the 'sentence' $Tsat(a,\sigma)$, but we'll leave doing so as an exercise. (In the case of arithmetic, this adds nothing, since every element of the domain has a name.)