

KRIPKE'S THEORY OF TRUTH

RICHARD G HECK, JR

1. INTRODUCTION

The purpose of this note is to give a simple, easily accessible proof of the existence of the minimal, and of various maximal, fixed points, under the Strong Kleene valuation scheme.¹ We begin by recording a few facts about unions and intersections which are used throughout, and which cannot be expected to be familiar to all students.

Definition. Let A be a set. Then $\cup A$ is the union of all the sets contained in A ; $\cap A$ is their intersection. That is, $x \in \cup A$ if, and only if, $\exists y(y \in A \wedge x \in y)$ and $x \in \cap A$ if, and only if, $\forall y(y \in A \rightarrow x \in y)$.

Proposition 1.1. *Let A be a set of sets and let y be some set. Then,*

- (1) *If y is a subset of some member of A , then y is a subset of $\cup A$. I.e., $\exists z(z \in A \wedge y \subseteq z) \rightarrow y \subseteq \cup A$. A fortiori, if $z \in A$, then $z \subseteq \cup A$.*
- (2) *If y is a superset of every member of A , then y is a superset of $\cup A$. I.e., $\forall z(z \in A \rightarrow z \subseteq y) \rightarrow \cap A \subseteq y$. A fortiori, if $z \in A$, then $\cap A \subseteq z$.*
- (3) *If y is a subset of every member of A , then y is a subset of $\cap A$. I.e., $\forall z(z \in A \rightarrow y \subseteq z) \rightarrow y \subseteq \cap A$.*
- (4) *If y is a superset of some member of A , then y is a superset of $\cap A$. I.e., $\exists z(z \in A \wedge z \subseteq y) \rightarrow \cap A \subseteq y$.*

Proof. We prove (1) and (2). For (1), suppose that y is a subset of z , which is a member of A . Then, since $z \in A$, $z \subseteq \cup A$. So $y \subseteq z \subseteq \cup A$, so $y \subseteq \cup A$. Since $z \subseteq z$, then of course $z \subseteq \cup A$.

For (2), suppose that y is a superset of every member of A . Let $x \in \cup A$. Then, by the definition of $\cup A$, there is some $w \in A$ such that $x \in w$. But $w \subseteq y$, so $x \in y$. And again, since $z \subseteq z$, of course $\cap A \subseteq z$. □

Exercise 1.2. Prove (3) and (4) from proposition 1.1.

¹I therefore have striven to keep the set-theoretic machinery to a minimum. No reference is made to ordinals: The only principle of set-theory not likely to be familiar to a beginner is Zorn's Lemma, which I assume, rather than prove (from choice). The method of proof used here therefore has certain disadvantages, as compared to Kripke's. Because ordinals are not used, and fixed-points are not constructed as the limits of sequences of theories, it is not possible to define a notion of level in these terms. Furthermore, while one can show that some sentences do not have truth-values in the minimal fixed point (are ungrounded), namely, by showing that they have different truth-values in some fixed points (see section 5), the technique obviously does not work for sentences that have intrinsic truth-values.

In developing the proof given here, I have drawn heavily upon work by Melvin Fitting [1]. The version of the proof given here is somewhat simpler than Fitting's, in large part because I have eschewed generality in favor of perspicuity. I independently discovered a proof dual to that given below in about 1990.

2. ZORN'S LEMMA

To state Zorn's Lemma, we need a few definitions.

Definition. Let C be a set of sets. C is said to be a *chain* iff $C \neq \emptyset$ and, for any two members x and y of C , one is a subset of the other. Thus, C is a chain only if $\forall x \forall y (x \in C \wedge y \in C \rightarrow x \subseteq y \vee y \subseteq x)$.

Definition. Let D be a set of sets. D is said to be *closed under unions of chains* iff the union of any chain which is a subset of D is an element of D . That is, D is closed under unions of chains iff: $\forall C [C \subseteq D \wedge C \text{ is a chain} \rightarrow \cup C \in D]$.

Definition. Let D be a set of sets. An element x of D is said to be a *maximal element of D* if it is the only member of D of which it is a subset, i.e., if there is no bigger set in D . Thus, x is maximal in D iff $\forall y (y \in D \wedge x \subseteq y \rightarrow x = y)$.

Lemma 2.1 (Zorn's Lemma). *Let D be a non-empty set which is closed under unions of chains and let $A \in D$. Then there is a maximal element of D of which A is a subset.*

Zorn's Lemma can be proven from the so-called Axiom of Choice (and the other axioms of set-theory), but we shall here simply take it as an axiom. (It is, in fact, equivalent to Choice, so this is reasonable.) Intuitively, the idea is this. Let A be an element of D and suppose, for *reductio*, that there is no superset of A that is a maximal element of D . If A had no proper superset, it would itself be maximal, so there is some bigger member of D of which A is a subset, call it A_1 . Similarly, A_1 must have a proper superset, call it A_2 ; which must in turn have a proper superset, A_3 ; and so on. Keep going in this way as long as you can; if you get through A_1, \dots, A_n, \dots , for all n , then take their union and keep going; etc. The sets A_i formed in this way will form a *chain* of elements of D , call it C : By construction, $A_i \subseteq A_{i+1}$.² Since D is closed under unions of chains, $\cup C$ in D . Moreover, $\cup C$ is a maximal element of D . For suppose not. Then $\cup C$ has some proper superset, call it y . But then $C \cup \{y\}$ is a chain of elements of D that is bigger than C , and in forming C we were supposed to keep going as long as we could. Contradiction.

3. KRIPKE'S CONSTRUCTION: INTERPRETATIONS

Definition. The T-rules are the following:

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

Our goal here is to prove that the T-rules are consistent with arithmetic (a similar construction will work, however, for any consistent first-order theory). The language of our theory is thus the language of arithmetic, augmented by the single, additional one-place predicate letter ' T '. We assume some fixed Gödel numbering of the sentences of our language, the details of which need

²It is to prove that this is possible that the Axiom of Choice is needed.

not concern us.³ A sentence of the form ' $T(\mathbf{n})$ '—here, we use ' \mathbf{n} ' as a name for the numeral denoting n , that is, for $S \dots S0$, with n occurrences of S —is intended to mean that the sentence with Gödel number n is true.⁴

We will construct an *interpretation* of our language that meets the following three desiderata:

- (1) all truths of arithmetic are true in that interpretation;
- (2) a sentence A is true under that interpretation if, and only if, $T(\ulcorner A \urcorner)$ is true under that interpretation;
- (3) a sentence A is false under that interpretation (i.e., its negation is true) if, and only if, $T(\ulcorner A \urcorner)$ is false under that interpretation (i.e., its negation is true).

Any such interpretation will be one in which all truths of arithmetic are true (by (1)) and in which the T-rules are *truth-preserving* (by (2) and (3)).

Exercise 3.1. Prove that no classical interpretation meets all of these conditions. Prove, that is, that, no matter what set we assign as extension of T , if we determine the truth-values of complex sentences in the usual way, the interpretation cannot satisfy all of (1)–(3).

What is an interpretation of this language to be like, then? The key idea is that we can allow a sentence of the form $T(\mathbf{n})$ not to have a truth-value at all. If so, of course, we cannot then use the usual truth-tables to determine the values negations, conjunctions, and the like, since the usual truth-tables assume that every sentence has a truth-value (either true or false).

Let us first discuss the logical constants. One can, in fact, proceed here in a variety of different ways. The simplest, however, is this. (This is called the Strong Kleene valuation scheme.) We want the truth-values of complex sentences to agree with those given by the truth-tables *if* all the constituent sentences themselves have truth-values; and we want to preserve certain basic intuitions, such as that a disjunction is true if either disjunct is true. So we make the following stipulations. First, we say that $\neg A$ is true if A is false; false, if A is true; and has no truth-value if A has no truth-value. Secondly, we say that $A \vee B$ is true, if either A or B is true; false, if both A and B are false; and has no truth-value, otherwise. We thus have, in effect, the following 'three-valued' truth-tables, where 'X' means 'has no truth-value':

\vee	T	X	F
T	T	T	T
X	T	X	X
F	T	X	F

The interpretations of ' \wedge ', ' \rightarrow ', and ' \equiv ' are then given by their usual definitions in terms of ' \neg ' and ' \vee '. So, $A \wedge B$ is true, if both A and B are true; false, if either A or B is false; and without truth-value, otherwise. $A \rightarrow B$ is true if either A is false or B is true; false, if A is true and B is false; and without truth-value, otherwise. And $A \equiv B$ is true if A and B have the same truth-value; false if they have different truth-values; and without truth-value if either A or B is without truth-value.

The quantifiers are then interpreted, as usual, as being infinite conjunction and infinite disjunction: A sentence of the form $\forall xA(x)$ will be true if $A(x)$ is true for every assignment of an object to ' x '; false, if $A(x)$ is false for some assignment to ' x '; and without truth-value otherwise. Similarly,

³For a simple presentation of the formal machinery needed here, see my "Formal Background for Theories of Truth".

⁴Henceforth, we shall frequently omit quotation marks, to improve readability.

$\exists xA(x)$ is true if $A(x)$ is true for some assignment to 'x'; false, if it is false for every assignment to 'x'; and without truth-value, otherwise.

How then are we to allow that a sentence of the form $T(\mathbf{n})$ might not have a truth-value? Kripke assigns an *extension* and an *anti-extension* to T and says that $T(\mathbf{n})$ is true if n is a member of the extension of T ; false, if it belongs to the anti-extension of T , and is without truth-value, otherwise. But there is a simpler method. We still assign just an extension to T , and we say that $T(\mathbf{n})$ is true if n is in the extension of T ; but we say that $T(\mathbf{n})$ is *false*—not if n is not in the extension of T , but—if n is the Gödel number of some sentence A and the Gödel number of the *negation* of A is in the extension of T : That is, $T(\ulcorner A \urcorner)$ is false iff $T(\ulcorner \neg A \urcorner)$ is true.⁵

We place one restriction upon the extensions assigned to T : They may not contain any numbers which are not the Gödel numbers of a sentence. Thus, if n is not the Gödel number of a sentence, then it will never be in the extension of T , so $T(\mathbf{n})$ will always be without truth-value. (This too is different from how Kripke does it: He insists that the Gödel numbers of non-sentences should be in the anti-extension.)

Given how we have interpreted T , the third desideratum on the interpretation we are seeking will follow from the second. These were:

2. a sentence A is true under a given interpretation if, and only if, $T(\ulcorner A \urcorner)$ is true under that interpretation;
3. a sentence A is false under the interpretation (i.e., its negation is true) if, and only if, $T(\ulcorner A \urcorner)$ is false under that interpretation (i.e., its negation is true).

Suppose that (2) holds. Let A be a sentence. Then $\neg A$ is true iff (by (2)) $T(\ulcorner \neg A \urcorner)$ is true iff (by the rules for interpreting T) $T(\ulcorner A \urcorner)$ is false iff (by the rules for negation) $\neg T(\ulcorner A \urcorner)$ is true. So $\neg A$ is true iff $\neg T(\ulcorner A \urcorner)$ is true, and we need not worry any more about (3).

We need only consider interpretations of the language which differ in what extension they assign to T . So we fix the rest of the interpretation. The *arithmetical* part of the language—consisting of '0', 'S', '+', and '×'—is to have its *intended* interpretation. (And '=' too is to be interpreted in the usual way). We take the universe of discourse, throughout, to comprise exactly the natural numbers 0, 1, etc. Desideratum (1) above is satisfied in any interpretation of this sort: For in any such interpretation, all the purely arithmetical sentences will be given their intended interpretation and so will be true in the interpretation just in case they are (really) true. Thus, the interpretations we shall be considering are fully determined by what extension is assigned to T .

Suppose we have an interpretation which assigns to T , as its extension, some set S . Some sentences come out true under this interpretation. Let $\mathbf{T}(S)$ be the set whose members are exactly the Gödel numbers of those sentences that come out true under this interpretation. Thus, $\mathbf{T}(x)$ is a function on sets.

Instead of saying, all the time, "A is true under the interpretation that assigns the set S as extension of ' T ' and is otherwise as said above", we shall say simply: A is true _{S} . Thus, $n \in \mathbf{T}(S)$ iff the sentence with Gödel number n is true _{S} . When we speak of a sentence's being true *simpliciter*, we mean sentences which really are true, such as '0 = 0'.

⁵I owe this trick to George Boolos.

Example. Consider the interpretation which assigns to T , as its extension, the empty set, \emptyset . What sentences will come out true under this interpretation? Well, certainly every true 'purely arithmetical' sentence—i.e., every true sentence not containing T —will be true_\emptyset , since the arithmetical part of the language is given its intended interpretation. There will be some sentences containing T that are true_\emptyset , as well, for example, $0 = 0 \vee T(\ulcorner 0 = 1 \urcorner)$, since $0 = 0$ is true_\emptyset , and so the disjunction of this sentence with any other sentence is true_\emptyset , as well. But no sentence of the form $T(\mathbf{n})$ is true_\emptyset , since the extension of T is empty. $\mathbf{T}(\emptyset)$ thus contains the Gödel numbers of all true purely arithmetical sentences, and some other numbers besides.

Example. Consider the interpretation which assigns to 'T', as its extension, the set of all Gödel numbers of sentences—call it $Sent$. What sentences come out true under this interpretation? Well, all true purely arithmetical sentences, again. And, if n is the Gödel number of a sentence, then $T(\mathbf{n})$ is true_{Sent} . So, in particular, both $T(\ulcorner 0 = 0 \urcorner)$ and $T(\ulcorner \neg 0 = 0 \urcorner)$ are true_{Sent} , which means that $T(\ulcorner 0 = 0 \urcorner)$ is also false_{Sent} —since $T(\ulcorner 0 = 0 \urcorner)$ is false_{Sent} iff $T(\ulcorner \neg 0 = 0 \urcorner)$ is true_{Sent} . So the interpretation that assigns the set of all Gödel numbers of sentences to S is, in an obvious sense, *inconsistent*: It makes some sentences come out both true and false.

Obviously, inconsistent interpretations are not going to do us much good. So we need to restrict our attention to consistent ones. Say that S is a *consistent set* if it does not contain both the Gödel number of A and that of ' $\neg A$ ', for any sentence A .

Definition. Con is the set of all consistent sets of Gödel numbers of sentences.

Lemma 3.2. *If S is a consistent set, then the interpretation that makes S the extension of T is a consistent interpretation. That is, if $S \in Con$, then no sentence is both true_S and false_S .*

Proof. The proof is by induction on the complexity of formulas. We show, first, that no atomic sentence can be both true_S and false_S ; and then we show that, if A and B are not both true_S and false_S , then neither are ' $\neg A$ ', ' $A \vee B$ ', ' $A \wedge B$ ', ' $A \rightarrow B$ ', and ' $A \equiv B$ '; and, that if no sentence of the form $A(\mathbf{n})$ is both true_S and false_S , then neither is $\forall xA(x)$ or $\exists xA(x)$.

If A is atomic, then it is either purely arithmetical or of the form $T(t)$, for some term t . If the former, it certainly cannot be both true_S and false_S , since it will have whatever truth-value it has in the intended interpretation of the language of arithmetic. If the latter, suppose that t denotes the number n . If $T(t)$ is true_S , $n \in S$; and if it is false_S , then the Gödel number of the negation of the sentence whose Gödel number is n is in S . But then S is not consistent, *contra* our supposition.

Suppose then that both A and B are not both true_S and false_S . If $\neg A$ is both true_S and false_S , then A itself must be both false_S and true_S . Similarly, if $A \wedge B$ is both true_S and false_S , both A and B must be true_S , and at least one of A and B must be false_S . Similarly for the conditional and biconditional.

So suppose that no sentence of the form $A(\mathbf{n})$ is both true_S and false_S . If $\forall xA(x)$ is both true_S and false_S , then, first, $A(x)$ must be true_S for every assignment to x , and so $A(\mathbf{n})$ must be true_S for every numeral \mathbf{n} . But ' $A(x)$ ' must also be false_S for at least one assignment to ' x ', so ' $A(\mathbf{n})$ ' must be false_S for *some* numeral \mathbf{n} , since every member of the universe of discourse is denoted by a numeral. But then some sentence ' $A(\mathbf{n})$ ' must be both true_S and false_S . Similarly for $\exists xA(x)$. \square

Note the trick used in the last paragraph: Because every object in the domain is the denotation of a numeral, we have that $\forall xA(x)$ is true iff $A(\mathbf{n})$ is true, for every numeral \mathbf{n} .

Corollary 3.3. *If S is consistent, so is $\mathbf{T}(S)$.*

Proof. Suppose $\mathbf{T}(S)$ is not consistent. Then, for some sentence A , $\mathbf{T}(S)$ must contain both the Gödel number of A and that of $\neg A$. Now, $\mathbf{T}(S)$ contains the Gödel numbers of sentences which are true_S. So both A and ' $\neg A$ ' must be true_S, i.e., A must be both true_S and false_S. But then, by 3.2, S is not consistent. \square

4. KRIPKE'S CONSTRUCTION: FIXED POINT MODELS

We are looking for an interpretation satisfying conditions (1)–(3), and to give such an interpretation we need only specify an appropriate set S as the extension of ' T '. Using our function $\mathbf{T}(x)$, it is easy to characterize the sets S that do the trick.

Definition. S is a *fixed-point* of $\mathbf{T}(x)$ iff S is a consistent set and $S = \mathbf{T}(S)$.

Proposition 4.1. *If S is a fixed point of $\mathbf{T}(x)$, then the interpretation that makes S the extension of T satisfies (1)–(3).*

Proof. As argued earlier, desideratum (3) is satisfied if (2) is. And (1) is satisfied by *any* interpretation of the sort we are considering. So we need only check that (2) is satisfied:

2. A is true_S if, and only if, $T(\ulcorner A \urcorner)$ is true_S.

But A is true_S iff (by the definition of \mathbf{T}) the Gödel number of A is in $\mathbf{T}(S)$ iff (since S is a fixed point) the Gödel number of A is in S iff (by the rules for determining the truth-value of $T(t)$) $T(\ulcorner A \urcorner)$ is true_S. \square

So all we now need to do is to show that there is a fixed point of $\mathbf{T}(x)$. We will in fact show two things: That $\mathbf{T}(x)$ has *maximal* fixed points and that $\mathbf{T}(x)$ has a *minimal* fixed point. By the definition given earlier, a fixed point is maximal if there is no other fixed point of which it is a (proper) subset. Such a fixed point will, as Kripke puts it, assign as many truth-values as is consistently possible. A fixed point is *minimal* if there is no other fixed point of which it is a superset. In fact, we shall see that there is a *unique* minimal fixed point that is a subset of every fixed point. Thus, the minimal fixed point is the one that contains the Gödel numbers of those sentences whose Gödel numbers must be in *any* fixed point.

Before giving the proofs, we need a couple preliminary results.

Lemma 4.2. *Let $S \subseteq R$. Then, if A is true_S, it is true_R. Moreover, if A is false_S, then it is false_R.*

Proof. By induction on the complexity of expressions. Certainly this holds if A is a purely arithmetical atomic sentence, since the truth of purely arithmetical sentences is not affected by what is assigned to T . Suppose that A is of the form $T(t)$ and that t denotes n . If $T(t)$ is true_S, then $n \in S$, but $S \subseteq R$, so $n \in R$, so $T(t)$ is true_R. And if $T(t)$ is false_S, then there must be some sentence B such that n is the Gödel number of B and the Gödel number of $\ulcorner \neg B \urcorner$, say m , is in S . But $S \subseteq R$, so $m \in R$, whence $T(t)$ is false_R, too.

Suppose now that A is logically complex, i.e., of one of the forms: $\neg B$, $B \vee C$, etc., $\forall xB(x)$, etc. We want to show that if B and C satisfy the lemma, then so do $\neg B$, $B \vee C$, etc., and that if all sentences of the form $B(\mathbf{n})$ do, then so do $\forall xB(x)$, etc.⁶ The proof is straightforward but tedious.

⁶Strictly speaking, of course, we can simply take the elided formulae to be defined in terms of the rest, so we do not really need to check them.

If A is $\neg B$ and A is true_S , then B must be false_S . But then, by the induction hypothesis, B must also be false_R , so $\ulcorner \neg B \urcorner$, i.e., A must be true_R . And if $\ulcorner \neg B \urcorner$ is false_S , then B is true_S , whence B is true_R , whence $\ulcorner \neg B \urcorner$ is false_R .

Similarly, if A is $\ulcorner B \vee C \urcorner$ and is true_S , at least one of B and C must be true_S ; but that sentence must also be true_R , by the induction hypothesis, whence $\ulcorner B \vee C \urcorner$ must also be true_R . And if $\ulcorner B \vee C \urcorner$ is false_S , both B and C must be false_S , whence they must both be false_R , whence $\ulcorner B \vee C \urcorner$ is false_R .

Similarly for the other propositional connectives.

If A is $\ulcorner \forall x B(x) \urcorner$ and is true_S , then $B(x)$ must be true_S , whatever may be assigned to ' x '. But then $B(\mathbf{n})$ must be true_S for every numeral \mathbf{n} ; so all such sentences must be true_R ; but then $B(x)$ is true_R , for every assignment to ' x ', and so $\ulcorner \forall x B(x) \urcorner$ is true_R , too. And if $\ulcorner \forall x B(x) \urcorner$ is false_S , $B(x)$ must be false_S for some assignment, say n , to ' x '; so $B(\mathbf{n})$ must be false_S and so false_R , whence $B(x)$ is false_R , when n is assigned to ' x ', and so $\ulcorner \forall x B(x) \urcorner$ is false_R .

Similarly for the existential quantifier. □

Corollary 4.3. *If $S \subseteq T$, then $\mathbf{T}(S) \subseteq \mathbf{T}(R)$. That is, $\mathbf{T}(x)$ is 'monotonic'.*

Proof. Suppose that $S \subseteq T$ and that $n \in \mathbf{T}(S)$. $\mathbf{T}(S)$ contains the Gödel numbers of sentences that are true_S . But any such sentence is, by 4.2, also true_R . So $n \in \mathbf{T}(R)$. □

This is the crucial fact for all that follows. As a close examination of the proofs will show, the only fact about \mathbf{T} upon which they rely is that \mathbf{T} is monotonic. Hence, our choice of the Strong Kleene valuation scheme is, in a certain sense, arbitrary: Other ways of handling the logical constants will also work, so long as the (analogous) operator \mathbf{T} defined in terms of them is monotonic.⁷

Lemma 4.4. *Con is closed under intersections and also under unions of chains. That is, if $B \subseteq \text{Con}$ and $B \neq \emptyset$, then $\cap B \in \text{Con}$ and, if B is a (non-empty) chain, then $\cup B \in \text{Con}$.*

Proof. Suppose $B \subseteq \text{Con}$, $B \neq \emptyset$. Then $\cap B$ is certainly a set of Gödel numbers of sentences. And it is consistent, since, if it contained both the Gödel number of some sentence A and that of its negation, every element of B would also contain these. But every element of B is consistent.

Suppose now that B is a chain. Then again, $\cup B$ is a set of Gödel numbers of sentences. Suppose, for *reductio*, that it is not consistent. Then there is some sentence A such that $\cup B$ contains both the Gödel number of A , say n , and that of $\ulcorner \neg A \urcorner$, say m . Since $n \in \cup B$, there must be some $D \in B$ such that $n \in D$. Similarly, for some $E \in B$, $m \in E$. Since B is a chain, either $D \subseteq E$ or $E \subseteq D$. If the former, then $n \in D \subseteq E$, so $n \in E$. But then E contains both n and m , i.e., both the Gödel number of A and that of $\ulcorner \neg A \urcorner$ and is therefore not consistent. Similarly, if $E \subseteq D$, then D is not consistent. So $\cup B$ is consistent. □

Sets S for which $S \subseteq \mathbf{T}(S)$ are of special interest. Such sets are often called 'sound', for this reason: Suppose that $S \not\subseteq \mathbf{T}(S)$. Then there is a sentence $A \in S$ such that $A \notin \mathbf{T}(S)$, i.e., such that A is not true_S . Thus, in the interpretation that treats S as the extension of the truth-predicate, $\mathbf{T}(\ulcorner A \urcorner)$ is true—that is, it is true_S —and yet A itself is *not* true in this interpretation—it is not true_S . On the other hand, if $S \subseteq \mathbf{T}(S)$, then any sentence $A \in S$ is true_S . So, in that sense, S is 'sound': The

⁷See my "Truth and Inductive Definability" for an examination of the sense in which this is so.

sentences that S 'says' are true are true _{S} . What the following theorem says is thus that every sound set can be extended to a fixed point.⁸

Theorem 4.5 (Fixed Point Theorem). Suppose that $S \in \text{Con}$ and that S is sound, i.e., that $S \subseteq \mathbf{T}(S)$. Then $\mathbf{T}(x)$ has a maximal fixed point of which S is a subset.

Proof. Let $B = \{X \in \text{Con} : X \subseteq \mathbf{T}(X)\}$. B is non-empty, since $S \in B$. We want to show that B is closed under unions of chains. So let $C \subseteq B$ be a (non-empty) chain. We must show that $\cup C \in \text{Con}$ and that $\cup C \subseteq \mathbf{T}(\cup C)$.

Now $B \subseteq \text{Con}$, so since $C \subseteq B$, $C \subseteq \text{Con}$. By 4.4, Con is closed under unions of chains. So $\cup C \in \text{Con}$.

To show that $\cup C \subseteq \mathbf{T}(\cup C)$, it suffices, by part (2) of proposition 1.1, to show that every member of C is a subset of $\mathbf{T}(\cup C)$. So suppose that $X \in C$. Then $X \subseteq \cup C$, trivially, so, since $\mathbf{T}(x)$ is monotonic, $\mathbf{T}(X) \subseteq \mathbf{T}(\cup C)$. Since $X \in C \subseteq B$, $X \in B$, so $X \subseteq \mathbf{T}(X)$. So $X \subseteq \mathbf{T}(X) \subseteq \mathbf{T}(\cup C)$, for any $X \in C$. So $\cup C \subseteq \mathbf{T}(\cup C)$.

So B is closed under unions of chains. By Zorn's Lemma, there is a maximal member of B of which S is a subset—call it M_S . We now show that M_S is a maximal fixed point of $\mathbf{T}(x)$. $M_S \in B$, so $M_S \subseteq \mathbf{T}(M_S)$. Since $\mathbf{T}(x)$ is monotonic, $\mathbf{T}(M_S) \subseteq \mathbf{T}(\mathbf{T}(M_S))$. And since M_S is consistent, $\mathbf{T}(M_S)$ is also consistent, by 3.2. Thus, $\mathbf{T}(M_S) \in \text{Con}$ and $\mathbf{T}(M_S) \subseteq B$. So $\mathbf{T}(M_S) \in B$. But M_S is a maximal element of B , so, since $M_S \subseteq \mathbf{T}(M_S)$, we have $M_S = \mathbf{T}(M_S)$, and so M_S is a fixed point of $\mathbf{T}(x)$.

Finally, M_S is a maximal fixed point. For let F be a fixed point other than M_S . So $F \in \text{Con}$ and $F = \mathbf{T}(F)$. But then, obviously, $F \subseteq \mathbf{T}(F)$, whence $F \in B$. But M_S is maximal in B , so F is not a superset of M_S . \square

Corollary 4.6. $\mathbf{T}(x)$ has at least one maximal fixed point.

Proof. Obviously, $\emptyset \subseteq \mathbf{T}(\emptyset)$. So, by 4.5, $\mathbf{T}(x)$ has a maximal fixed point of which \emptyset is a subset. \square

Theorem 4.7 (Kripke). $\mathbf{T}(x)$ has a unique minimal fixed point.

Proof. Let $B = \{X \in \text{Con} : \mathbf{T}(X) \subseteq X\}$. Let M be a maximal fixed point of $\mathbf{T}(x)$. Then $M \in \text{Con}$ and $M = \mathbf{T}(M)$. So B is non-empty.

We now show that $\cap B$ is a fixed point of $\mathbf{T}(x)$. Certainly, $\cap B \in \text{Con}$, since Con is closed under intersections, by 4.4.

To show that $\cap B = \mathbf{T}(\cap B)$, we show both that $\mathbf{T}(\cap B) \subseteq \cap B$ and that $\cap B \subseteq \mathbf{T}(\cap B)$.

To prove the former, it is enough to show, by part (3) of 1.1, that $\mathbf{T}(\cap B)$ is a subset of every member of B . So assume $X \in B$. Then $\cap B \subseteq X$, so since $\mathbf{T}(x)$ is monotonic, $\mathbf{T}(\cap B) \subseteq \mathbf{T}(X)$. But since $X \in B$, $\mathbf{T}(X) \subseteq X$, so $\mathbf{T}(\cap B) \subseteq X$.

To prove the latter, it is enough to show that $\mathbf{T}(\cap B) \in B$. For then certainly $\cap B \subseteq \mathbf{T}(\cap B)$, since $\cap B$ is a subset of every member of B , by proposition 1.1. Now, $\cap B \in \text{Con}$, so $\mathbf{T}(\cap B) \in \text{Con}$, by 3.3. Since, by the argument of the last paragraph, $\mathbf{T}(\cap B) \subseteq \cap B$. But $\mathbf{T}(x)$ is monotonic, so $\mathbf{T}(\mathbf{T}(\cap B)) \subseteq \mathbf{T}(\cap B)$. Thus, $\mathbf{T}(\cap B) \in \text{Con}$ and $\mathbf{T}(\mathbf{T}(\cap B)) \subseteq \mathbf{T}(\cap B)$, so $\mathbf{T}(\cap B) \in B$. So $\cap B = \mathbf{T}(\cap B)$, and $\cap B$ is a fixed point.

⁸For this reason, Michael Kremer calls sound sets 'fixable'.

Let F be any fixed point of $\mathbf{T}(x)$. Then $F \in \text{Con}$ and $\mathbf{T}(F) = F$, so certainly $\mathbf{T}(F) \subseteq F$. Hence $F \in B$, whence $\cap B \subseteq F$. So $\cap B$ is minimal. \square

Note that the minimal fixed point is *not* the intersection of all *maximal* fixed points. In fact, it is not obvious that the intersection of all maximal fixed points is a fixed point at all, or even a sound set.

5. APPLICATIONS

We know from Tarski's Theorem that not all instances of ' $A \equiv T(\ulcorner A \urcorner)$ ' can be true. But we can say a bit more about their status in Kripke's construction.

Remark 5.1. No instance of ' $A \equiv T(\ulcorner A \urcorner)$ ' is false in any fixed point.

Proof. Suppose ' $A \equiv T(\ulcorner A \urcorner)$ ' is false $_S$, where S is a fixed point. Then either A is true $_S$ and $T(\ulcorner A \urcorner)$ is false $_S$ or conversely. But if A is true $_S$, then, since fixed points satisfy the T-rules, $T(\ulcorner A \urcorner)$ must also be true $_S$, so S is not consistent. Similarly, if $T(\ulcorner A \urcorner)$ is true $_S$, then A is true $_S$. \square

Definition. A sentence is *grounded* if it has a truth-value in the interpretation that assigns as extension of T the minimal fixed point: That is, if either its Gödel number or that of its negation is in the minimal fixed point.

Definition. A sentence is *paradoxical* if it has no truth-value in any interpretation that assigns as extension of ' T ' some fixed point: That is, if neither its Gödel number nor that of its negation is in any fixed point.

Exercise 5.2. Show that a sentence is grounded if it has the same truth-value in every fixed point and, equivalently, has *some* truth-value in every fixed point.

It is important to note that the diagonal lemma *does not hold in its classical form in this framework*. According to Gödel's version of the diagonal lemma, and that usually employed, for any formula $A(x)$, there is a sentence G such that

$$\vdash G \equiv A(\ulcorner G \urcorner)$$

But, in the logic in which we are working, it might be that both G and $A(\ulcorner G \urcorner)$ are without truth-value, whence $G \equiv A(\ulcorner G \urcorner)$ will be without truth-value, too. The diagonal lemma does, however, hold in a modified form.

Lemma 5.3 (Modified Diagonal Lemma). *There is a Gödel numbering of the formulae of the language of arithmetic which is such that, if $A(x)$ is any formula, then there is a term t such that:*

$$Q \vdash t = \ulcorner A(t) \urcorner$$

Moreover, there is a sentence G , namely, $A(t)$, such that:

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

where the double turnstile means that the inference is valid in both directions.

We shall not prove this version of the diagonal lemma here.⁹

Consider now the formula $\neg T(x)$. There is, by the modified diagonal lemma, a sentence λ such that:

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

Thus, λ is a liar sentence.

Remark 5.4. λ is paradoxical: Neither the Gödel number of λ nor that of its negation belongs to any fixed point.

Proof. Suppose that F is a fixed point and that the Gödel number of λ is in F . Then $T(\ulcorner \lambda \urcorner)$ is true_F . But then, since F is a fixed point, λ must be true_F , as well. But by construction, λ implies $\neg T(\ulcorner \lambda \urcorner)$, so $\neg T(\ulcorner \lambda \urcorner)$ is also true_F . But then F is not consistent, by 3.2. Similarly, if $\ulcorner \neg \lambda \urcorner$ is in F , then $T(\ulcorner \neg \lambda \urcorner)$ is true_F , so $\neg T(\ulcorner \lambda \urcorner)$ is false_F ; but $\neg T(\ulcorner \lambda \urcorner)$ implies λ , and again F is not consistent. \square

Remark 5.5. $\lambda \equiv T(\ulcorner \lambda \urcorner)$ is paradoxical: It has no truth-value in any fixed point.

Proof. It is without truth-value whenever λ is. \square

Now consider the formula $T(x)$. There is, by the modified diagonal lemma, a term τ such that $Q \vdash \tau = \ulcorner T(\tau) \urcorner$. So τ is a truth-teller.

Remark 5.6. There are fixed points that contain $\ulcorner T(\tau) \urcorner$ and there fixed points that contain $\ulcorner \neg T(\tau) \urcorner$.

Proof. Consider the set $\{\tau\}$. This set is obviously consistent. Moreover, $\{\tau\} \subseteq \mathbf{T}(\{\tau\})$. For $\tau \in \mathbf{T}(\{\tau\})$ just in case the sentence with Gödel number τ is $\text{true}_{\{\tau\}}$. But since, obviously, $\tau \in \{\tau\}$, we have that $T(\tau)$ is $\text{true}_{\{\tau\}}$; but $T(\tau)$ is the sentence Gödel number τ . So $\{\tau\}$ is both sound and consistent. By 4.5, then, there is a (maximal) fixed point M of which $\{\tau\}$ is a subset, that is, of which τ is a member.

Consider now the set $\{\text{neg}(\tau)\}$, where $\text{neg}(\tau)$ is the Gödel number of the negation of the sentence with Gödel number τ , i.e., that if $\neg T(\tau)$. This set too is obviously consistent. Moreover, $\{\text{neg}(\tau)\} \subseteq \mathbf{T}(\{\text{neg}(\tau)\})$. For $\text{neg}(\tau) \in \mathbf{T}(\{\text{neg}(\tau)\})$ just in case the sentence with Gödel number $\text{neg}(\tau)$ is $\text{true}_{\{\text{neg}(\tau)\}}$. But since, obviously, $\text{neg}(\tau) \in \{\text{neg}(\tau)\}$, we have that $T(\text{neg}(\tau))$ is $\text{true}_{\{\text{neg}(\tau)\}}$; so $T(\tau)$ is $\text{false}_{\{\text{neg}(\tau)\}}$, whence $\neg T(\tau)$ is $\text{true}_{\{\text{neg}(\tau)\}}$; but, as noted, $\text{neg}(\tau) = \ulcorner \neg T(\tau) \urcorner$, so $\text{neg}(\tau)$ is indeed $\text{true}_{\{\text{neg}(\tau)\}}$. So $\{\text{neg}(\tau)\}$ is both sound and consistent. By 4.5, there is a (maximal) fixed point M of which it is a subset, that is, of which $\text{neg}(\tau)$ is a member. \square

Remark 5.7. τ is ungrounded: Neither its Gödel number nor that of its negation is in the minimal fixed point.

Proof. The minimal fixed point is a subset of every maximal fixed point. If τ were in the minimal fixed point, then, it would have to be in every fixed point. But, as we just saw, there are fixed points that contain $\text{neg}(\tau)$. Such a fixed point cannot contain $\ulcorner \tau \urcorner$, on pain of its not being consistent. Similarly, $\text{neg}(\tau)$ cannot be in the minimal fixed point either. \square

⁹Such a result for the case of a language more expressive than that of arithmetic—in particular, for the language of primitive recursive arithmetic—is originally due to Jeroslow [3]. The present form is mentioned by Kripke [4] and is proven in detail in my “Self-reference and the Languages of Arithmetic” [2]. See also [REF Visser].

Remark 5.8. $T(\ulcorner T(\ulcorner \tau \urcorner) \urcorner) \equiv T(\ulcorner \tau \urcorner)$ is ungrounded. But there is a fixed point F such that it is true_F .

Proof. The mentioned sentence will be without truth-value whenever $T(\ulcorner \tau \urcorner)$ is, and that is without truth-value in the minimal fixed point. But there is a fixed point F in which $T(\ulcorner \tau \urcorner)$ is true_F , whence $T(\ulcorner T(\ulcorner \tau \urcorner) \urcorner)$ is true_F , and then so is $T(\ulcorner T(\ulcorner \tau \urcorner) \urcorner) \equiv T(\ulcorner \tau \urcorner)$. \square

Remark 5.9 (Adapted from Vann McGee [5]). Let A be any sentence. Then there is a sentence G such that, for every fixed point F , if $G \equiv T(\ulcorner G \urcorner)$ is true_F , then A is true_F .

Proof. Consider the formula $T(x) \equiv A$. By the modified diagonal lemma, there is a sentence G such that:

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

Now, let F be a fixed point and suppose that $G \equiv T(\ulcorner G \urcorner)$ is true_F . Then both G and $T(\ulcorner G \urcorner)$ have a truth-value $_F$, and they have the same one.

Suppose that G is true_F . By the T-rules, $T(\ulcorner G \urcorner)$ is also true_F . But then, by the first of the displayed rules, $T(\ulcorner G \urcorner) \equiv A$ must also be true_F , and so A must be true_F .

Suppose, then, that $\neg G$ is true_F . By the T-rules, $T(\ulcorner \neg G \urcorner)$ is also true_F ; so $T(\ulcorner G \urcorner)$ is false_F . But then, by the second of the displayed rules, $\neg(T(\ulcorner G \urcorner) \equiv A)$ is true_F , so $T(\ulcorner G \urcorner) \equiv A$ is false_F . So $T(\ulcorner G \urcorner)$ and A must both have truth-values $_F$, and these must be different. Since $T(\ulcorner G \urcorner)$ is false_F , A must be true_F .

Either way, then, A is true_F , as claimed. \square

Thus, every sentence follows from some instance of the T-schema, if the T-rules hold (and if the biconditional works as it does in the Strong Kleene scheme).

REFERENCES

- [1] Melvin Fitting. Notes on the mathematical aspects of kripke's theory of truth. *Notre Dame Journal of Formal Logic*, 27:75–88, 1986.
- [2] Richard G. Heck. Self-reference and the languages of arithmetic. *Philosophia Mathematica*, 15:1–29, 2007.
- [3] R. G. Jeroslow. Redundancies in the Hilbert-Bernays derivability conditions for Gödel's second incompleteness theorem. *Journal of Symbolic Logic*, 38:359–67, 1973.
- [4] Saul Kripke. Outline of a theory of truth. *Journal of Philosophy*, 72:690–716, 1975.
- [5] Vann McGee. Maximal consistent sets of instances of tarski's schema (t). *Journal of Philosophical Logic*, 21:235–41, 1992.