1. INTRODUCTION

The purpose of this note is to give an accessible introduction to the construction of the minimal fixed point of Kripke’s theory of truth [5], under the Strong Kleene valuation scheme.\(^1\) A background in basic set theory, together with some mathematical sophistication, should be enough for the reader to understand this note.

2. THREE-VALUED INTERPRETATIONS

**Definition.** The T-rules are the following:

\[
\begin{align*}
A & \vdash T(\langle A \rangle) \\
\neg A & \vdash \neg T(\langle A \rangle) \\
T(\langle A \rangle) & \vdash A \\
\neg T(\langle A \rangle) & \vdash \neg A
\end{align*}
\]

Our goal here is to prove that the T-rules are consistent with arithmetic.\(^2\) 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 not concern us.\(^3\) A sentence of the form ‘\(T(n)\)’—here, we use \(\langle n \rangle\) as a name of the numeral denoting \(n\) (that is, of: \(S\ldots S0\))—is intended to mean that the sentence with Gödel number \(n\) is true.\(^4\)

---

\(^1\)The underlying results are not presented in full generality: That is not needed for the applications made here. For that, see Moschovakis [8], Kechris and Moschovakis, [4], and McGee [6, ch. 5]. Useful papers here, besides Kripke’s original, include Feferman [1].

\(^2\)A similar construction will work, however, for any consistent first-order theory.

\(^3\)For a simple presentation of the formal machinery needed here, see my “Formal Background for Theories of Truth”.

\(^4\)Henceforth, we shall frequently omit quotation marks, to improve readability.
We'll assume throughout that the arithmetical part of the language is interpreted in the usual way.\footnote{This also is inessential: We can fix an arbitrary interpretation of the language, and the rest of the construction will then go through unchanged.}

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

\textbf{K1:} All truths of arithmetic are true in that interpretation.

\textbf{K2:} A sentence $A$ is true under that interpretation if, and only if, $T(\langle A \rangle)$ is true under that interpretation.

\textbf{K3:} A sentence $A$ is false under that interpretation (i.e., its negation is true) if, and only if, $T(\langle A \rangle)$ 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 (K1), and in which the T-rules are truth-preserving, by (K2) and (K3).

What is an interpretation of this language to be like, then? The key idea is that we can allow a sentence of the form $T(\langle n \rangle)$ not to have a truth-value at all. If so, of course, we cannot then use the usual truth-tables to determine the values of negations, conjunctions, and the like, since the usual truth-tables assume that every sentence has a truth-value (either true or false). Note also that the usual rule of conditional proof will have to be abandoned, that is, the rule that tells us that if $\Sigma, A \models B$, then $\Sigma \models A \rightarrow B$. Otherwise, since $T(\langle A \rangle) \models A$, we will have $\models T(\langle A \rangle) \rightarrow A$; since $A \models T(\langle A \rangle)$, then $\models A \rightarrow T(A)$; but then $\models T(A) \equiv A$, which looks worrying.

Let us first discuss the logical constants. One can actually proceed here in a variety of different ways. The simplest, however, which is known as the ‘strong (or Kleene) valuation scheme’, is this. 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 $A \lor B$ 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-table, where ‘$X$’ means ‘has no truth-value’:

<table>
<thead>
<tr>
<th>$\lor$</th>
<th>$T$</th>
<th>$X$</th>
<th>$F$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
<td>$T$</td>
</tr>
<tr>
<td>$X$</td>
<td>$T$</td>
<td>$X$</td>
<td>$X$</td>
</tr>
<tr>
<td>$F$</td>
<td>$T$</td>
<td>$X$</td>
<td>$F$</td>
</tr>
</tbody>
</table>

The interpretations of ‘$\land$’, ‘$\rightarrow$’, and ‘$\equiv$’ are then given by their usual definitions in terms of ‘$\neg$’ and ‘$\lor$’. So, $A \land B$ 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 of them is without truth-value.

The quantifiers are then interpreted, as usual, as being (in effect) infinite conjunction and infinite disjunction: A sentence of the form \( \forall x A(x) \) 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, \( \exists x A(x) \) 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.

Since we are interested here in the language of arithmetic, we can simplify our treatment of the quantifiers. Every object in the domain is a natural number, and every natural number has a standard name: the numeral that denotes it. So \( A(x) \) is true for every assignment of an object to \( 'x' \) iff \( A(n) \) is true for every \( n \); and \( A(x) \) is false for some assignment to \( 'x' \) iff \( A(n) \) is false for some \( n \). In the present context, then, we can characterize the truth of quantified sentences in terms of the truth of their instances rather than having to talk, more generally, about satisfaction. The techniques used below can be applied to the case of satisfaction as well as truth, but doing so introduces complications that we will not consider here.

We can now ensure that the first desideratum (K1) will be satisfied simply by interpreting the arithmetical part of the language in the usual way. So the domain will be the set of natural numbers; \( '0' \) will denote zero; \( 'S' \), succession; \( '+' \), addition; \( 'x' \), multiplication; and \( '<' \), less than. No matter what the interpretation of the predicate-letter \( T \), all the purely arithmetical sentences will be given their intended interpretation and so will be true in our interpretation just in case they are (really) true.\(^6\) Thus, we need only consider, henceforth, what a particular interpretation does with \( T \).

How are we to allow that a sentence of the form \( T(n) \) might not have a truth-value? One standard way to treat predicates in a three-valued setting is to take them to have an extension and an anti-extension, where the extension is the set of objects of which the predicate is true, and the anti-extension is the set of objects of which it is false.\(^7\) This is how Kripke proceeds. But we may simplify a bit: We still assign just an extension to \( T \), and we say that \( T(n) \) is true if \( n \) is in the extension of \( T \); but we say

\(^6\)This depends, of course, upon the fact that the three-valued truth-tables agree with the classical tables when everything has a truth-value.

\(^7\)More generally still, in a many-valued setting, one can take the semantic value of a predicate to be a function from objects to truth-values. In the present case, one could adopt that strategy or one could take there to be just the two truth-values, true and false, and allow the function to be partial.
that \( T(\bar{n}) \) 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, \( T(\neg A) \) is false iff \( T(\neg \neg A) \) is true.\(^8\)

We will write \( \neg n \) to mean: the Gödel number of the negation of the sentence whose Gödel number is \( n \).

We place one restriction upon the extensions assigned to \( T \): They may not contain any number which is not the Gödel number 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 \( T(\bar{n}) \) will always be without truth-value.\(^9\)

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

**K2:** A sentence \( A \) is true under a given interpretation if, and only if, \( T(\neg A) \) is true under that interpretation.

**K3:** A sentence \( A \) is false under the interpretation (i.e., its negation is true) if, and only if, \( T(\neg A) \) is false under that interpretation (i.e., its negation is true).

Suppose that (K2) holds, and let \( A \) be a sentence. Then \( \neg A \) is true iff (by (K2)) \( T(\neg A) \) is true iff (by the rules for interpreting \( T \)) \( T(\neg A) \) is false iff (by the rules for negation) \( \neg T(\neg A) \) is true. So \( \neg A \) is true iff \( \neg T(\neg A) \) is true, and we need not worry any more about (K3).

### 3. The Kripke Jump

Suppose we have an interpretation which assigns to \( T \), as its extension, some set \( S \). Some sentences come out true under this interpretation. Let \( T(S) \) be the set whose members are exactly the Gödel numbers of those sentences that come out true under this interpretation. Thus, \( T(x) \) is an operator (function) on sets. We may call it the ‘Kripke jump’ operation.

Instead of saying, all the time, “\( A \) is true under the interpretation that assigns the set \( S \) as extension of \( T \) and is otherwise the standard interpretation of the language of arithmetic”, we shall say simply: \( A \) is true\(_S\). Thus, \( n \in T(S) \) iff the sentence with Gödel number \( n \) is true\(_S\). When we speak of a sentence’s being true simpliciter, we mean it really is true, in the way that ‘0 = 0’ is true.

**Example.** Consider the interpretation which assigns to \( T \), as its extension, the empty set, \( \emptyset \). What sentences will come out true under this interpretation? Which sentences are true\(_\emptyset\)? 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 also be some sentences containing \( T \) that are true\(_\emptyset\), as well, for example, \( 0 = 0 \lor T(\neg 0 = 1) \), since \( 0 = 0 \) is

\(^8\)I owe this trick to George Boolos.

\(^9\)This is different from how Kripke does it: He insists that the Gödel numbers of non-sentences should be in the anti-extension.
true∅, and so the disjunction of this sentence with any other sentence is true∅, as well. But no sentence of the form T(⌜n⌝) is true∅, since the extension of T is empty. T(⌜∅⌝) thus contains the Gödel numbers of all true purely arithmetical sentences and some other sentences that follow logically from those.

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(⌜n⌝) is trueSent. So, in particular, both T(⌜0 = 0⌝) and T(⌜¬0 = 0⌝) are trueSent, which means that T(⌜0 = 0⌝) is also falseSent—since T(⌜0 = 0⌝) is falseSent iff T(⌜¬0 = 0⌝) is trueSent. So the interpretation that assigns the set of all Gödel numbers of sentences to T 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.

Definition 3.1.
(i) S is a consistent set if it does not contain both the Gödel number of A and that of ¬A, for any sentence A; i.e., if n ∈ S, then ¬n /∈ S.
(ii) 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 ∈ Con, then no sentence is both trueS and falseS.

Proof. The proof is by induction on the complexity of formulas. We show, first, that no atomic sentence can be both trueS and falseS; and then we show that, if A and B are not both trueS and falseS, then neither are ¬A, A ∨ B, A ∧ B, A → B, and A ≡ B; and, that if no sentence of the form A(⌜n⌝) is both trueS and falseS, then neither is ∀xA(x) or ∃xA(x).

The atomic sentences are either purely arithmetical or of the form T(t), for some term t. But no purely arithmetical sentence can be both trueS and falseS, since those have whatever truth-value they have in the intended interpretation of the language of arithmetic. So consider a sentence of the form T(t), and suppose that t denotes the number n. If T(t) is trueS, then n ∈ S; if it is also falseS, then ¬n ∈ S. But then S /∈ Con, contra our supposition.

Suppose then that both A and B are not both trueS and falseS. If ¬A is both trueS and falseS, then A itself must be both falseS and trueS. Similarly, if A ∧ B is both trueS and falseS, both A and B must be trueS, and at least one of A and B must be falseS.

Finally, suppose that no sentence of the form A(⌜n⌝) is both trueS and falseS. If ∀xA(x) is both trueS and falseS, then, first, A(⌜n⌝) must be trueS.
for every \( n \). But \( A(\overline{n}) \) must also be \( \text{false}_S \) for some \( n \). But then some sentence \( A(\overline{n}) \) must be both \( \text{true}_S \) and \( \text{false}_S \).

The other cases are left as Exercise 8.2.

\[ \square \]

**Corollary 3.3.** If \( S \) is a consistent set, then so is \( T(S) \). That is: If \( S \in \text{Con} \), then \( T(S) \in \text{Con} \).

**Proof.** Suppose \( T(S) \) is not consistent. Then, for some sentence \( A \), \( T(S) \) must contain both the Gödel number of \( A \) and that of \( \neg A \). Now, \( T(S) \) contains the Gödel numbers of sentences which are \( \text{true}_S \). So both \( A \) and \( \neg A \) must be \( \text{true}_S \), i.e., \( A \) must be both \( \text{true}_S \) and \( \text{false}_S \). 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 (K1)–(K3), and to give such an interpretation we need only specify an appropriate set \( S \) as the extension of \( T \). Using our function \( T(S) \), it is easy to characterize the sets \( S \) that will do the trick.

**Definition 4.1.** \( S \) is a fixed-point of \( T(x) \) iff \( S \in \text{Con} \) and \( S = T(S) \).

**Proposition 4.2.** If \( S \) is a fixed point of \( T(x) \), then the interpretation that makes \( S \) the extension of \( T \) satisfies (K1)–(K3).

**Proof.** As argued earlier, desideratum (K3) is satisfied if (K2) is. And (K1) is satisfied by any interpretation of the sort we are considering. So we need only check that (K2) is satisfied.

We may re-write (K2) in the form:

**K2:** \( A \) is \( \text{true}_S \) if, and only if, \( T(\overline{\neg A}) \) is \( \text{true}_S \).

But \( A \) is \( \text{true}_S \) iff (by the definition of \( T \)) \( \overline{\neg A} \in T(S) \) iff (since \( S \) is a fixed point) \( \overline{\neg A} \in S \) iff (by the rules for determining the truth-value of sentences of the form \( T(t) \)) \( \overline{T(\overline{\neg A})} \) is \( \text{true}_S \). \[ \square \]

So all we now need to do is to show that there is a fixed point of \( T(x) \). We will in fact show that \( T(x) \) has lots of fixed points, including what is known as the (unique) minimal fixed point. This is a fixed point that is contained in every fixed point: It contains the Gödel numbers of those sentences whose Gödel numbers must be in any fixed point.

The idea behind the proof is this. Suppose we start with an interpretation that makes the extension of \( T \) just be the empty set. Then, as we saw above, there are some sentences that will be counted as true under this interpretation, namely, those in \( T(\emptyset) \). That will not be a fixed point,\(^\text{10}\) but we can repeat the procedure: Let the extension of \( T \) be \( T(\emptyset) \) and consider what sentences come out true in that interpretation. That will give us \( T(T(\emptyset)) \). This once again will not be a fixed point. In fact, even running through all the finite iterations will not get us to a fixed point.

\(^{10}\) Since, e.g., \( T(\overline{T(\overline{0 = 0})}) \) is \( \text{true}_{T(\emptyset)} \) but \( T(\overline{T(\overline{0 = 0})}) \notin T(\emptyset) \).
point. (See Exercise 8.3.) So take the union of everything we have so far and keep going. And keep on doing that: applying T repeatedly and then taking unions when you run out of opportunities to do that. The hope is that, if we keep going in this way for long enough, we will eventually reach a fixed point.

The really crucial point is that iterating the Kripke jump in this way gives us a non-decreasing sequence of sets of true sentences. That is: \( \emptyset \subseteq T(\emptyset) \subseteq T(T(\emptyset)) \subseteq T(T(T(\emptyset))), \) and so forth. More generally, we have the following.

**Lemma 4.3.** If \( S \subseteq R, \) then \( T(S) \subseteq T(R). \) That is, \( T(x) \) is ‘monotonic’.

This is the essential fact about \( T \) for all that follows. As a close examination of the proofs will show, the only fact about \( T \) upon which they rely is that \( T \) 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 \( T \) defined in terms of them is monotonic.\(^{11}\)

It will suffice to prove Lemma 4.3 to establish the following.

**Lemma 4.4.** Let \( S \subseteq R. \) Then, if \( A \) is true\(_S\), it is true\(_R\). Moreover, if \( A \) is false\(_S\), then it is false\(_R\).

For then suppose \( S \subseteq R \) and \( n \in T(S) \). Then \( n \) is the Gödel number of some sentence \( A \) that is true\(_S\); but then, by Lemma 4.4, \( A \) is also true\(_R\). So \( n \in T(R) \). We need to deal simultaneously with truth and falsity because negation, in particular, ‘converts’ between the two. (This is a common feature of proofs in this area.)

**Proof.** By induction on the complexity of expressions. Certainly this holds for purely arithmetical atomic sentences, since the truth of purely arithmetical sentences is not affected by the extension of \( T \). So consider a sentence of the form \( T(t) \), and suppose 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 \( \neg n \in S \). But \( S \subseteq R \), so \( \neg n \in R \), whence \( T(t) \) is false\(_R\), too.

Suppose \( \neg B \) is true\(_S\). Then \( B \) must be false\(_S\). So, by the induction hypothesis, \( B \) must also be false\(_R\), so \( \neg B \) must be true\(_R\). And if \( \neg B \) is false\(_S\), then \( B \) is true\(_S\), whence \( B \) is true\(_R\), whence \( \neg B \) is false\(_R\).

Similarly, if \( B \lor C \) 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 \( B \lor C \) must also be true\(_R\). And if \( B \lor C \) is false\(_S\), both \( B \) and \( C \) must be false\(_S\), whence they must both be false\(_R\), whence \( B \lor C \) is false\(_R\).

If \( \forall x B(x) \) is true\(_S\), then \( B(\bar{n}) \) must be true\(_S\) for every \( n \); so all such sentences must be true\(_R\); but then \( \forall x B(x) \) is true\(_R\), too. And if \( \forall x B(x) \)

\(^{11}\)These include the ‘weak (or Fregean) valuation scheme’, which takes a formula to have no truth-value whenever any one of its components does, and the ‘supervaluational’ scheme, which I shall not try to describe here.
is false, \( B(\bar{n}) \) must be false for some \( n \), but then \( B(\bar{n}) \) is also false, whence \( \forall x B(x) \) is false.

The other sentential connectives and the existential quantifier are left as Exercise Exercise 8.4.

Monotonicity allows us to show, as said above, that iterated application of the Kripke jump operator produces a non-decreasing set of sets of sentences. Since \( \emptyset \subseteq T(\emptyset) \), trivially, we also have \( T(\emptyset) \subseteq T(T(\emptyset)) \), by monotonicity; hence also \( T(T(\emptyset)) \subseteq T(T(T(\emptyset))) \), again by monotonicity, and so forth. Indeed, this sort of phenomenon will occur whenever we ‘start out’ with a set \( S \) for which \( S \subseteq T(S) \). Then, by monotonicity, \( T(S) \subseteq T(T(S)) \); hence, by monotonicity again, \( T(T(S)) \subseteq T(T(T(S))) \), and so forth. Such sets are thus of special interest.

**Definition 4.5.** A set \( S \) is sound if \( S \subseteq T(S) \).

Here’s the reason for the terminology. Suppose that \( S \not\subseteq T(S) \). Then there is a sentence \( A \) whose Gödel number is in \( S \), in which case \( T(\langle A \rangle) \) is true, and yet since \( \langle A \rangle \not\in T(S) \), \( A \) itself is not true. So the interpretation that takes \( S \) to be the extension of \( T \) ‘says’, as it were, that \( T(\langle A \rangle) \) is true when \( A \) is not true. On the other hand, if \( S \subseteq T(S) \), then, if \( \langle A \rangle \subseteq S \), so that \( T(\langle A \rangle) \) is true, then \( A \) also is true. So the sentences that \( S \) ‘says’ are true are indeed true.

**Theorem 4.6 (Fixed Point Theorem).** Let \( S \) be a consistent, sound set. Then there is a fixed point \( F_S \) of \( T(x) \) that (i) contains \( S \) and (ii) is itself consistent.

We’ll prove Theorem 4.6 in the sections that follow. Obviously, it entails the existence of fixed points.

**Corollary 4.7.**

(i) \( T(x) \) has at least one consistent fixed point, namely, \( F_\emptyset \).

(ii) Moreover, this is the minimal fixed point of \( T \), in the sense that it is contained in every fixed point.

**Proof.** Part (i) is easy: Obviously, \( \emptyset \subseteq T(\emptyset) \), so \( \emptyset \) is sound; hence, by Theorem 4.6, there is a fixed point \( F_\emptyset \) (that contains \( \emptyset \)).

The proof of part (ii) requires transfinite induction, which we’ll introduce shortly.

\[ \Box \]

5. **Ordinals**

We shall need some basic facts about ordinal numbers.\(^{12}\) 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.

\[ ^{12} \text{The theory of ordinal numbers can be developed in set theory, which treats ordinal numbers as sets of a certain kind. See any decent textbook on set theory.} \]
A sequence is well-ordered if it is linearly ordered—transitive, irreflexive, and connected—and if it satisfies 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:

(Transitive) $\forall x \forall y \forall z (x < y \land y < z \rightarrow x < z)$

(Irreflexive) $\forall x \forall y (x < y \rightarrow \neg y < x)$

(Connected) $\forall x \forall y (x < y \lor x = y \lor y < x)$

(Least) $\forall R (R \subseteq S \land R \neq \emptyset \rightarrow \exists y \in R \forall z \in R (y \leq z))$

The natural numbers 1, 2, 3, and so forth can be construed as ordinals, pronounced ‘first’, ‘second’, ‘third’, and so forth. They mark the corresponding positions in the following sequence:

$1, 2, 3, \ldots$

There are also infinite ordinals. Consider the following sequence:

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

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 $\omega^{th}$ position: The first infinite ordinal is $\omega$ (omega). The next position, occupied by Brutus, is the $(\omega + 1)^{st}$ position; the next, occupied by Claudius, the $(\omega + 2)^{nd}$. And so on.

Consider now the following sequence:

$-1, -2, -3, \ldots, 0, 1, 2, 3, \ldots, \text{Caesar, Brutus, Claudius}$

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

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

(There are $\omega$ copies of $\omega$, in effect, followed by Caesar.) And so on through $\omega^2 + \omega, \omega^2 + 2\omega, \omega^2 + 3\omega, \omega(\omega^2) = \omega^3, \omega^4, \omega^\omega,$ 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 $\alpha$ is a successor ordinal if $\alpha = \beta + 1$, for some ordinal $\beta$. Otherwise, $\alpha$ is a limit.
Examples of limit ordinals are 0 (which is the only finite limit), $\omega$, and $\omega^2$.

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 then occupies. The answer is $\omega_1$, the first uncountable ordinal. And the ordinals do not stop there. But we shall not need 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 5.1.** 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.

We shall at various points need to appeal to certain elementary facts about ordinal arithmetic, for example, that the ordinals satisfy the law of trichotomy: $\alpha < \beta \lor \alpha = \beta \lor \beta < \alpha$. We'll call attention to these facts as they are needed, but not prove them here. They can all be proven in set theory, given appropriate definitions.

What we shall really need to know 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: First, one shows that 0 is $F$ (the basis case); then one shows that, if a given number $n$ is $F$, then so is $n + 1$ (the induction step). An equivalent method of proof, sometimes called ‘strong induction’, replaces that sort of induction step with this one: If all numbers less than $n$ are $F$, then so is $n$. One uses this kind of induction, for example, in proving that every number has a unique prime

---

13Strictly speaking, there is no ordinal 0, since there is no zero position of a sequence. But we allow 0 for convenience.

14That is, a set is countable if there is a one-one function from the set of natural numbers, $\mathbb{N}$, onto that set.

15This is not trivial, but it can be proven in set theory, given an appropriate definition of the ordinals.

16One does need to be careful here, since ordinal arithmetic is different from cardinal arithmetic. For example, ordinal addition is not commutative: $2 + \omega$ is just $\omega$ again, since the thing that occupies the $\omega$th place after the $2$nd thing is just the $\omega$th thing.

17Strong induction follows from the usual sort of induction. Suppose we want to show that $\forall x F x$ and we know that $\forall n[\forall z < n (F z) \rightarrow F n]$. Then consider $\forall x < n (F x)$ and argue by induction in the usual way.
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., 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.

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.

6. Monotone Operators

We can straightforwardly generalize the notion of monotonicity that was introduced earlier.

**Definition 6.1.** Let \( A \) be some set of sets that is the domain and range of some operator (function) on sets \( \Phi \). So \( \Phi : A \rightarrow A \). Then:

(i) \( \Phi \) is **monotonic** if, whenever \( S \subseteq T \), \( \Phi(S) \subseteq \Phi(T) \).

(ii) \( S \) is a **fixed point** of \( \Phi \) if \( \Phi(S) = S \).

**Theorem 6.2 (Fixed-point Theorem).** Let \( \Phi : A \rightarrow A \) be monotonic. Then \( \Phi \) has a fixed point and, indeed, a unique minimal fixed point.

We'll prove the theorem for countable sets \( A \), but the proof readily generalizes to larger sets.

We have already described the basic idea behind the proof. We start with some sound set \( S \) and start applying \( \Phi \), forming the sequence \( S \subseteq \Phi(S) \subseteq \Phi(\Phi(S)) \subseteq \cdots \) and continue in this way. The various inclusions hold because \( S \) is sound and \( \Phi \) is monotonic. 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 \( \Phi \), taking unions at any limits we reach. Eventually, we will reach a fixed point, or so the proof we are about to give shows.

First, we need rigorously to define the sequence of sets just described.

**Definition 6.3.** The function \( I^{\alpha,S}_\Phi \) is defined by transfinite recursion thus:

\[
I^{0,S}_\Phi = S \\
I^{\alpha,S}_\Phi = \Phi(\bigcup_{\xi<\alpha} I^{\xi,S}_\Phi), \text{ for } \alpha \neq 0
\]
Here, $\bigcup_{\xi<\alpha} I_\Phi^{\xi,S}$ is the union of all the sets $I_\Phi^{\xi,S}$ for $\xi<\alpha$.

The ‘steps’ of this construction are conventionally known as its stages, and they are indexed by the ordinals. So we may speak of the first stage, the $\omega$th stage, and so forth.

We now need to show that this sequence is non-decreasing, in the sense that later members of the sequence contain earlier members. This will hold only for the analogue of sound sets, as we defined them earlier.

**Definition 6.4.** $S$ is sound for $\Phi$ if $S \subseteq \Phi(S)$.

**Lemma 6.5.** If $\Phi$ is monotonic and $S$ is sound for $\Phi$, then $S \subseteq I_\Phi^{\alpha,S}$, for all $\alpha$.

**Proof.** If $\alpha = 0$, then of course $S \subseteq \Phi(S) = I_\Phi^{0,S}$, by soundness. If, on the other hand, $\alpha \neq 0$, then note that $I_\Phi^{0,S} \subseteq \bigcup_{\xi<\alpha} I_\Phi^{\xi,S}$, since $I_\Phi^{0,S}$ just is one of the $I_\Phi^{\xi,S}$, and the union of any bunch of sets contains each of those sets. So

$$S \subseteq \Phi(S), \text{ by soundness}$$

$$= \Phi(I_\Phi^{0,S}), \text{ by definition of } I_\Phi^{0,S}$$

$$\subseteq \Phi\left(\bigcup_{\xi<\alpha} I_\Phi^{\xi,S}\right), \text{ by monotonicity}$$

$$= I_\Phi^{\alpha,S}, \text{ by definition of } I_\Phi^{\alpha,S}$$

□

**Lemma 6.6.** If $\Phi$ is monotonic and $S$ is sound for $\Phi$, then $\forall \beta \forall \alpha (\beta < \alpha \rightarrow I_\Phi^{\beta,S} \subseteq I_\Phi^{\alpha,S})$. In particular, $I_\Phi^{\alpha} \subseteq I_\Phi^{\alpha+1}$.

**Proof.** If $\alpha = 0$, then this is trivial, since no $\beta < 0$. So suppose $\alpha \neq 0$ and suppose $\beta < \alpha$. We need to show that $I_\Phi^{\beta,S} \subseteq I_\Phi^{\alpha,S}$. There are two cases: Either $\beta = 0$ or not. If so, then $I_\Phi^{0,S} = S \subseteq I_\Phi^{\alpha,S}$, by Lemma 6.5. If $\beta \neq 0$, then note first that $\bigcup_{\xi<\beta} I_\Phi^{\xi,S} \subseteq \bigcup_{\xi<\alpha} I_\Phi^{\xi,S}$, since the latter union includes all the sets the former one does. But then:

$$I_\Phi^{\beta,S} = \Phi\left(\bigcup_{\xi<\beta} I_\Phi^{\xi,S}\right), \text{ by definition of } I_\Phi^{\beta,S}$$

$$\subseteq \Phi\left(\bigcup_{\xi<\alpha} I_\Phi^{\xi,S}\right), \text{ by monotonicity}$$

$$= I_\Phi^{\alpha,S}, \text{ by definition of } I_\Phi^{\alpha,S}$$

□

**Corollary 6.7.** $I_\Phi^{\alpha+1,S} = \Phi(I_\Phi^{\alpha,S})$

---

18Because $<$ on the ordinals is transitive: If $\xi < \beta < \alpha$, then $\xi < \alpha$. 
Proof: By definition, \( I_{\Phi}^{\alpha+1,S} = \Phi(\bigcup_{\xi<\alpha+1} I_{\Phi}^{\xi,S}) \). I claim that \( \bigcup_{\xi<\alpha+1} I_{\Phi}^{\xi,S} = I_{\Phi}^\alpha \). Now \( \xi < \alpha + 1 \) iff \( \xi = \alpha \) or \( \xi < \alpha \). So \( \bigcup_{\xi<\alpha+1} I_{\Phi}^{\xi,S} = I_{\Phi}^\alpha \cup \bigcup_{\xi<\alpha} I_{\Phi}^{\xi,S} \). But if \( \xi < \alpha \), then \( I_{\Phi}^{\xi,S} \subseteq I_{\Phi}^{\alpha,S} \), by Lemma 6.6. So \( \bigcup_{\xi<\alpha} I_{\Phi}^{\xi,S} \subseteq I_{\Phi}^{\alpha} \).\(^{19}\) Hence, \( I_{\Phi}^\alpha \cup \bigcup_{\xi<\alpha} I_{\Phi}^{\xi,S} = I_{\Phi}^{\alpha,S} \).

What that means is that the second clause in Definition 6.3, which was formulated so as to apply to any ordinal, reduces in the case of successors to the much simpler condition: \( I_{\Phi}^{\alpha+1,S} = \Phi(I_{\Phi}^{\alpha,S}) \). So \( I_{\Phi}^{\xi,S} \) acts as explained above: We are building a sequence by applying \( \Phi \) repeatedly, and then taking unions at limits.

We can now restate Theorem 6.2 in somewhat sharper form.

**Theorem 6.8** (Fixed-point Theorem, Refined). Let \( \Phi : A \to A \) be monotonic, where \( A \) is countable, and suppose that \( S \) is sound for \( \Phi \). Then there is a (least) countable ordinal \( \alpha \) such that fixed point \( F_S \) of \( \Phi \) such that \( S \subseteq F_S \).

Moreover, \( F_{\emptyset} \) is the (unique) minimal fixed point of \( \Phi \), in the sense that it is a subset of every fixed point.

**Proof.** Suppose that for no countable \( \alpha \) do we have \( \Phi(I_{\Phi}^{\alpha,S}) = I_{\Phi}^{\alpha,S} \). Then for every \( \alpha \), \( I_{\Phi}^{\alpha,S} \subseteq \Phi(I_{\Phi}^{\alpha,S}) = I_{\Phi}^{\alpha+1,S} \). Hence, for each countable \( \alpha \), we can find \( \varphi(\alpha) \in I_{\Phi}^{\alpha+1,S} \setminus I_{\Phi}^{\alpha,S} \).\(^{20}\) We want now to show that \( \varphi \) is one-one.

Suppose \( \alpha \neq \beta \); so either \( \alpha < \beta \) or conversely; without loss of generality, suppose \( \alpha < \beta \). Then \( \varphi(\beta) \in I_{\Phi}^{\beta+1,S} \setminus I_{\Phi}^{\beta,S} \), whence \( \varphi(\beta) \notin I_{\Phi}^{\beta,S} \), whereas \( \varphi(\alpha) \in I_{\Phi}^{\alpha+1,S} \), by definition of \( \varphi \). Since \( \alpha < \beta \), \( \alpha + 1 \leq \beta \), whence \( I_{\Phi}^{\alpha+1,S} \subseteq I_{\Phi}^{\beta,S} \), by Lemma 6.6, so \( \varphi(\beta) \notin I_{\Phi}^{\alpha+1,S} \); hence \( \varphi(\alpha) \neq \varphi(\beta) \).

So \( \varphi \) is a one-one function from all the countable ordinals into \( A \). But \( A \) is countable, so that contradicts Fact 5.1.\(^{21}\) Thus, there is a countable ordinal \( \alpha \) such that \( \Phi(I_{\Phi}^{\alpha,S}) = I_{\Phi}^{\alpha,S} \). Henceforth let \( \alpha \) be the least such ordinal.\(^{22}\) So \( I_{\Phi}^{\alpha,S} \) is a fixed point of \( \Phi \), which we may now dub \( F_S \). Moreover, by Lemma 6.6, again, \( S = I_{\Phi}^{0,S} \subseteq I_{\Phi}^{\alpha,S} = F_S \), so that completes the proof of the first part of the theorem.

Since \( \emptyset \) is trivially sound for \( \Phi \), there is a least countable ordinal \( \alpha \) such that \( F_{\emptyset} = I_{\Phi}^{\alpha,\emptyset} \) is a fixed point of \( \Phi \). We need to show that \( I_{\Phi}^{\alpha,\emptyset} \) is

---

\(^{19}\)If one set in a union contains all the others, then the union is just that set.

\(^{20}\)The notation means: The members of \( I_{\Phi}^{\alpha+1,S} \) that are not in \( I_{\Phi}^{\alpha,S} \). (The Axiom of Choice is actually needed for this step.)

\(^{21}\)That is, if we never reached a fixed point at a countable ordinal, then \( I_{\Phi}^{\alpha,S} \) would have kept growing through all the countable ordinals: At least one new element would have been added at each such ordinal. So if we were to collect all of these together, they would form a set the same size as the set of countable ordinals. So they would form an uncountable subset of \( A \). But \( A \) is countable, so that is impossible.

\(^{22}\)Since the ordinals are themselves well-ordered by \( \langle \), they satisfy a form of the least number principle: If there are any ordinals that have some property \( F \), then there is a (unique) least one.
minimal, i.e., that it is contained in every fixed point of $\Phi$. So let $F$ be any fixed point of $\Phi$. We show by transfinite induction that $I^{\beta,0}_\Phi \subseteq F$, for all $\beta$, from which it follows that $I^{\alpha,0}_\Phi \subseteq F$.

**Basis:** Obviously, if $\beta = 0$, then $I^{0,0}_\Phi = \emptyset \subseteq F$.

**Induction step:** Suppose $I^{\xi,0}_\Phi \subseteq F$, for all $\xi < \beta$. We need to show that $I^{\beta,0}_\Phi \subseteq F$. Now, a union of subsets of $F$ must itself be a subset of $F$, so

$$
I^{\beta,0}_\Phi = \Phi(\bigcup_{\xi<\beta} I^{\xi,0}_\Phi), \text{ by definition of } I^{\beta,0}_\Phi
$$

$$
\subseteq \Phi(F), \text{ by monotonicity}
$$

$$
= F, \text{ since } F \text{ is a fixed point}
$$

As claimed. 

Most of Theorem 4.6 now follows immediately from Theorem 6.8. In particular, there is a least countable ordinal $\alpha$ such that $I^{\alpha,0}_T$ is the (unique) least minimal fixed point of $T(x)$, which also proves the second part of Corollary 4.7. There is, however, one thing we still need to check, namely, that the fixed points we construct in this way are consistent.

**Proposition 6.9.** Suppose $S$ is both sound and consistent. Then $I^{\alpha,S}_T$ is consistent, for all $\alpha$. In particular, the fixed point $F_S$ of $T$ is consistent.

**Proof.** By transfinite induction. The basis is trivial, since $I^{0,S}_T = S$, which we have assumed to be consistent.

For the induction step, suppose that $I^{\beta,S}_T$ is consistent for all $\beta < \alpha$. We need to show that $I^{\alpha,S}_T = T(\bigcup_{\xi<\alpha} I^{\xi,S}_T)$ is consistent. It will suffice to show that $\bigcup_{\xi<\alpha} I^{\xi,S}_T$ is consistent since, by Corollary 3.3, $T(\bigcup_{\xi<\alpha} I^{\xi,S}_T)$ is then also consistent.

So suppose that both $\neg A^\gamma$ and $\neg \neg A^\gamma$ are in $\bigcup_{\xi<\alpha} I^{\xi,S}_T$. Then, for some $\beta < \alpha$, $\neg A^\gamma \in I^{\beta,S}_T$, and for some $\gamma < \alpha$, $\neg \neg A^\gamma \in I^{\gamma,S}_T$. Now either $\beta < \gamma$ or $\beta = \gamma$ or $\gamma < \beta$. If $\beta = \gamma$, however, then $\neg \neg A^\gamma \in I^{\beta,S}_T$, so $I^{\beta,S}_T$ is not consistent, contrary to the induction hypothesis. And if $\beta < \gamma$, then $I^{\beta,S}_T \subseteq I^{\gamma,S}_T$, by Lemma 6.6; so $\neg A^\gamma \in I^{\gamma,S}_T$, and $I^{\gamma,S}_T$ is not consistent, again contrary to the induction hypothesis. Similarly, if $\gamma < \beta$, then $\neg \neg A^\gamma \in I^{\beta,S}_T$, and $I^{\beta,S}_T$ is not consistent. 

### 7. Applications

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

---

23 If an object is an element of the union of some sets, then it is an element of one of those sets.
Proposition 7.1. No instance of \( A \equiv T(\langle A \rangle) \) is false in any fixed point.

Proof. Let \( F \) be a fixed point. Suppose \( A \) is true\(_F\). Then, since fixed points satisfy the T-rules, \( T(\langle A \rangle) \) must also be true\(_F\), in which case \( A \equiv T(\langle A \rangle) \) is true\(_F\). Similarly, if \( A \) is false\(_F\), then \( T(\langle A \rangle) \) must also be false\(_F\), and \( A \equiv T(\langle A \rangle) \) is again true\(_F\). And if \( A \) is neither true nor false, then \( A \equiv T(\langle A \rangle) \) is neither true nor false. So it is never false. \( \square \)

Definition 7.2. A sentence is grounded if it has a truth-value in the interpretation that make the 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.

Proposition 7.3. Let \( A \) be a grounded sentence. Then \( A \equiv T(\langle A \rangle) \) is true in the minimal fixed point.

Proof. Since \( A \) is grounded, it has a truth-value in the minimal fixed point. But then it must be true, by Proposition 7.1. \( \square \)

It is important to note that the diagonal lemma does not hold in its usual form in this framework. According to Gödel's version of the diagonal lemma, for any formula \( A(x) \), there is a sentence \( G_A \) such that

\[
\vdash G_A \equiv A(\langle G_A \rangle)
\]

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

Lemma 7.4 (Strong 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:

\[
Q \vdash t = \langle A(t) \rangle
\]

Moreover, there is a sentence \( G_A \), namely, \( A(t) \), such that:

\[
G_A \models A(\langle G_A \rangle)
\]

\[
A(\langle G_A \rangle) \not\models G_A
\]

\[
\neg G_A \not\models A(\langle G_A \rangle)
\]

\[
\neg A(\langle G_A \rangle) \not\models \neg G_A
\]

We shall not prove this version of the diagonal lemma here.\(^{24}\) We will, however, use it.

\(^{24}\)Such 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 [5] and is proven in detail in my “Self-reference and the Languages of Arithmetic” [2].
Consider the formula \( \neg T(x) \). There is, by the modified diagonal lemma, a term \( \lambda \) such that \( Q \vdash \lambda = \neg T(\lambda) \), and there is a sentence \( \Lambda \)—namely, \( \neg T(\lambda) \), whose Gödel number is, to emphasize, \( \lambda \)—such that:

\[
\begin{align*}
\Lambda & \models \neg T(\Lambda) \\
\neg T(\Lambda) & \not\models \Lambda \\
T(\Lambda) & \not\models \neg \Lambda \\
\neg \Lambda & \models T(\Lambda)
\end{align*}
\]

Thus, \( \Lambda \) is a liar sentence.

**Definition 7.5.** A sentence is **paradoxical** if neither its Gödel number nor that of its negation is in any fixed point.

**Proposition 7.6.** \( \Lambda \) is paradoxical.

**Proof.** Suppose that \( F \) is a fixed point and that \( \lambda \) is in \( F \). Then \( T(\lambda) \) is true_\( F \). But then, since \( F \) is a fixed point, \( \lambda \in T(F) = F \), so the sentence with Gödel number \( \lambda \) is true_\( F \), as well; i.e., \( \neg T(\lambda) \) is true_\( F \). But then \( F \) is not consistent, by 3.2. Similarly, if \( \neg \lambda \) is in \( F \), then \( T(\neg \lambda) \) is true_\( F \). But since \( F \) is a fixed point, the sentence with Gödel number \( \neg \lambda \) is true_\( F \), as well; i.e., \( \neg (\neg T(\lambda)) \) is true_\( F \). But then \( \neg T(\lambda) \) is false_\( F \), and so \( F \) is once again not consistent. \( \square \)

Why can’t we construct a fixed point containing \( \lambda \) by starting our construction with \( \{ \lambda \} \)? Answer:

**Fact 7.7.** \( \{ \lambda \} \) is not sound. Indeed, no consistent set containing \( \lambda \) is sound.

**Proof.** Let \( S \) be a consistent set, and suppose \( \lambda \in S \). Then \( T(\lambda) \) is true_\( S \), so \( \neg T(\lambda) \not\in T(S) \). But then \( \neg T(\lambda) \not\models \Lambda \), and so \( S \) is not sound. \( \square \)

**Proposition 7.8.** Let \( A \) be a paradoxical sentence. Then \( A \equiv T(\neg A) \) is also paradoxical. In particular, \( \lambda \equiv T(\neg \lambda) \) is paradoxical.

**Proof.** Suppose \( A \equiv T(\neg A) \) has a truth-value in some fixed point \( F \). By Proposition 7.1, it must be true_\( F \). So \( A \) and \( T(\neg A) \) must have the same truth-value_\( F \). But then \( A \) has a truth-value in \( F \) and so is not paradoxical. \( \square \)

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

**Proposition 7.9.** There are fixed points that contain \( T(\tau) \) and there fixed points that contain \( \neg T(\tau) \).

**Proof.** We prove just the first part and leave the second as Exercise 8.6. Consider the set \( \{ \tau \} \). This set is obviously consistent. Moreover, \( \{ \tau \} \subseteq T(\{ \tau \}) \). For \( \tau \in T(\{ \tau \}) \) just in case the sentence with Gödel
number \( \tau \) is true_{\{\tau\}}. But since \( \tau \in \{\tau\} \), we have that \( T(\tau) \) is true_{\{\tau\}}; but \( T(\tau) \) is the sentence with Gödel number \( \tau \), so \( \tau \in T(\{\tau\}) \). So \( \{\tau\} \) is both sound and consistent. So there is a (minimal) fixed point of which \( \{\tau\} \) is a subset, that is, of which \( \tau \) is a member. □

**Proposition 7.10.** \( T(\tau) \) 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 fixed point. If \( \lceil T(\tau) \rceil \) were in the minimal fixed point, then, it would have to be in every fixed point. But there are fixed points that contain \( \lceil T(\tau) \rceil \). Such a fixed point cannot contain \( \lceil \neg T(\tau) \rceil \), since fixed points are consistent; hence, \( \lceil \neg T(\tau) \rceil \) cannot be in the minimal fixed point. Similarly, \( \lceil T(\tau) \rceil \) cannot be in the minimal fixed point. □

**Proposition 7.11.** \( T(\lceil \neg T(\tau) \rceil) \equiv T(\tau) \) is ungrounded. But there is a fixed point \( F \) such that it is true_{\tau_F}.

**Proof.** The mentioned sentence will be without truth-value whenever \( T(\tau) \) is, and \( T(\tau) \) is without truth-value in the minimal fixed point. But there is a fixed point \( F \) in which \( T(\tau) \) is true_{\tau_F}, whence \( T(\lceil \neg T(\tau) \rceil) \equiv T(\tau) \) is also true_{\tau_F} and so is \( T(\lceil T(\tau) \rceil) \equiv T(\tau) \). □

**Proposition 7.12** (Adapted from Vann McGee [7]). Let \( A \) be any sentence. Then there is a sentence \( G \) such that, for every fixed point \( F \), if \( G \equiv T(\lceil G \rceil) \) is true_{\tau_F}, then \( A \) is true_{\tau_F}.

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

\[
\begin{align*}
G &\vdash T(\lceil G \rceil) \equiv A \\
\neg G &\vdash \neg[T(\lceil G \rceil) \equiv A]
\end{align*}
\]

Now, let \( F \) be a fixed point and suppose that \( G \equiv T(\lceil G \rceil) \) is true_{\tau_F}. Then both \( G \) and \( T(\lceil G \rceil) \) have a truth-value_{\tau_F}, and they have the same one. Suppose that \( G \) is true_{\tau_F}. By the T-rules, \( T(\lceil G \rceil) \) is also true_{\tau_F}. But then, by the first of the displayed implications, \( T(\lceil G \rceil) \equiv A \) must also be true_{\tau_F}, and so \( A \) must be true_{\tau_F}. Suppose, then, that \( \neg G \) is true_{\tau_F}. By the T-rules, \( T(\lceil \neg G \rceil) \) is also true_{\tau_F}; so \( T(\lceil G \rceil) \) is false_{\tau_F}. But then, by the second of the displayed implications, \( \neg[T(\lceil G \rceil) \equiv A] \) is true_{\tau_F}, so \( T(\lceil G \rceil) \equiv A \) is false_{\tau_F}. So \( T(\lceil G \rceil) \) and \( A \) must both have truth-values_{\tau_F}, and these must be different. Since \( T(\lceil G \rceil) \) is false_{\tau_F}, \( A \) must be true_{\tau_F}.

Either way, then, \( A \) is true_{\tau_F}, as claimed.

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).
8. Exercises

Exercise 8.1. Prove that no classical interpretation meets all three of the conditions mentioned on page 2. Prove, that is, that, no matter what set we assign as extension of $T$, if we take $T(t)$ to be false if (as usual) the denotation of $t$ is not in the extension of $T$, then the interpretation cannot satisfy all of (1)–(3). (Hint: This is a corollary of Tarski’s Theorem, i.e., of the liar paradox.)

Exercise 8.2. Do the cases of disjunction and existential quantification that we omitted in the proof of 3.2.

Exercise 8.3. Show that none of $T(\emptyset)$, $T(T(\emptyset))$, $T(T(T(\emptyset)))$, etc, is a fixed point. (See Note 10 for a hint.) Optional extra: Show that the union of all those (what we have at stage $\omega$ of the iterative construction) is not a fixed point either. (Hint: Consider the sentence that says that all of the sentences you previously constructed are true. The trick is to see that this can be expressed in the language of arithmetic.)

Exercise 8.4. Do the cases of conjunction and existential quantification that we omitted in the proof of Lemma 4.4.

Exercise 8.5. (Optional) We might alternatively have defined $I_{\Phi}^{\alpha,S}$ as follows:

\[
I_{\Phi}^{0,S} = S \\
I_{\Phi}^{\alpha+1,S} = \Phi(I_{\Phi}^{\alpha,S}) \\
I_{\Phi}^{\alpha,S} = \bigcup_{\beta < \alpha} I_{\Phi}^{\beta,S}
\]

Show that we can then still prove Lemma 6.5 and Lemma 6.6 and so can still prove Theorem 6.8. In what ways is this definition better or worse than the one we gave? (Hint: The proofs of the two lemmas using the original definition did not use transfinite induction.)

Exercise 8.6. Show that there fixed points that contain $\neg T(\tau)$.

Exercise 8.7. The natural way to try to formalize “The Liar is neither true nor false” would be $\neg T(\lambda) \land \neg T(\neg \lambda)$. Show that this is actually paradoxical.

Exercise 8.8. (Optional) Diagonalize on $T(x) \lor \neg T(x)$ to get a term $\kappa$ such that $Q \vdash \kappa = \neg T(\kappa) \lor \neg T(\neg \kappa)$. What can we say about the sentence $T(\kappa) \lor \neg T(\kappa)$? What if we diagonalize on $T(x) \land \neg T(x)$?

Exercise 8.9. Suppose that there were some formula $U(x)$ of the language of arithmetic plus $T$ such that $U(\bar{n})$ is true, no matter what the interpretation of $T$, iff $n$ is the Gödel number of an ungrounded sentence. (Note that we are making no assumptions about what happens when $\bar{n}$ is
the Gödel number of a grounded sentence.) Show that there would then be a term \( \gamma \) denoting the sentence \( U(\gamma) \lor \neg T(\gamma) \) and that we could then reproduce the liar paradox in Kripke's theory. (Hint: Suppose this sentence is grounded, and show that that is impossible. So it is ungrounded. But then it is true.)

So we cannot express the notion of ungroundedness in Kripke's theory; that is, there is no way truly to say from within Kripke's theory that the liar is ungrounded. (This is the 'ghost of the Tarski hierarchy'.)

REFERENCES