Casper Storm Hansen

Grounded Ungroundedness

Abstract. A modification of Kripke's theory of truth is proposed and it is shown how this modification solves some of the problems of expressive weakness in Kripke's theory. This is accomplished by letting truth values be grounded in facts about other sentences' ungroundedness.

Keywords: Truth, paradoxes, grounding, Kripke's theory of truth, expressive strength

1. Introduction

Kripke's well-known theory of truth [13] has some (also well-known) problems with regard to semantic openness and inadequate modeling of some intuitively unproblematic uses of the truth predicate. I will present a modification of the theory that solves some of these problems. But first (section 2) it is argued that the basic version of Kripke's theory is on the right track if we are looking for an explication of the correspondence theory of truth, because the correspondence relation is a grounding relation. The modification is done in an attempt to stay true to these basic ideas behind Kripke's construction and just take them a step further by extending the range of facts that truth values can be grounded in to include facts about sentences being ungrounded. Thereby some of the problems of expressive weakness in Krikpe's own theory are solved.

I will assume familiarity with Kripke's theory[1] and only describe some of its problems of expressive weakness, one here and one in section 9.

While the Liar sentence

The Liar: The Liar is false

[1]Kripke presents different versions of his theory. When I write as if there were *one* theory which is Kripke's, I mean the basic version, i.e. the smallest fixed point version using Kleene's strong three-valued logic presented on pages 700–705 of [13]. In the absence of a story about how to "reach them from below", the larger fixed points cannot reasonably be seen as explications of the correspondence theory as the true and false sentences are not grounded in the way explained in the following. I reserve discussion of supervaluation for another paper.
A distinction has been made between an external and an internal version of Kripke's theory [8]. I take Kripke's theory to be the external one and see the difference as being a discrepancy between what is the case according to the theory and what can be expressed.

is undefined (neither true nor false) by Kripke's theory, this is not a fact that can be expressed in the object-language itself; the sentence "The Liar is undefined" can not be formalised, much less made true. For the language only contains choice negation that takes truth to falsity, falsity to truth, and undefinedness to undefinedness, and not exclusion negation that takes undefinedness to truth. So the negation of the truth predicate is, in effect, a falsity predicate (for any sentence s, $\neg T(\text{`}s\text{'})$ is true iff $T(\text{`}s\text{'})$ is false iff s is false) and does not mean what "not true" intuitively means. Hence, there is no way to "build" an undefinedness predicate out of the truth predicate and the connectives. This problem can be characterised as one of semantic openness with regard to individual semantic facts. This is how Kripke avoids the revenge problem: By making sure that the revenge liar for the theory, "This sentence is either false or undefined", cannot be formulated. His method for avoiding inconsistency is essentially the same as Tarski's, namely through expressive weakness, or what Kripke [13, 714] himself refers to as the "ghost of the Tarski hierarchy". In Tarski's hierarchy a sentence cannot be about its own truth, and in Kripke's a sentence cannot be about its own, nor any other sentence's, undefinedness. But the undefinedness of the Liar is a fact, and like other facts it should be capable of being the ground for true sentences. In this paper I will let it do so. Consistency is instead retained by carefully discerning what grounded correspondence truth can amount to.

2. An interpretation of Kripke's theory

It seems quite reasonable to take Kripke's theory as an explication of the correspondence theory of truth. That this is so is best seen, I think, from the book metaphor which Beall [2, chapter 1] uses to introduce the theory, so let me recapitulate it here.

Imagine a world initially consisting only of non-semantic facts. In this world, there is a writer with two very large books. They carry the titles *The True* and *The False*. In the beginning they are empty, but the writer sets out to fill them so that they accurately reflect their titles. In the first book, he records every fact in the world, and in the second, he records every state of affairs that fails to obtain in the world. For instance, he writes "Snow is white" in the first book and "Snow is green" in the second. After having done so, he realises that his work is not complete. For now there are more facts than when he started, and those facts are perfectly capable of grounding new truths and falsehoods. By writing in the books, he has added facts to the world, namely facts about what is written in the books, and he did not include these facts in *The True*, nor did he include non-obtaining facts about the books in *The False*. So in each book, he puts the heading "Chapter 1"

over what he has written so far and starts writing the more comprehensive second chapters of each book. Chapter 2 of *The True* is a complete record of all facts about the world outside the books as well as about chapter 1 of each of the books. He uses the predicate "is true" to mean "is a sentence written in *The True*" and similarly the predicate "is false" to mean "is a sentence written in *The False*". So "'Snow is white' is true" and "'Snow is green' is false" both appear in this chapter. Because "Snow is green" is in *The False*, it is determined that this sentence will never be in *The True*, no matter how many chapters are written, so the writer can put "'Snow is green' is true" in chapter 2 of *The False*. After having written the two new chapters, there are again new facts, so the writer also compiles increasingly comprehensive chapters 3, 4, 5, etc.

It is quite obvious that this writer adheres strictly to the correspondence theory of truth in writing his books. A sentence is considered true (written in *The True*) just when the state of affairs described by the sentence obtains – outside the books or in the books depending on what kind of state of affairs the sentence is about. He is creating a well-founded correspondence relation, and that is exactly what I take the idea of grounding to be all about (at least in the domain of truths and truth makers).

According to the correspondence theory of truth, a sentence is true if it corresponds to or represents a fact. For something to represent something else, the represented must in some sense be logical prior to the representation. So when not only the representation but also the represented is a sentence, i.e. when a sentence is about sentences, an order of dependency appears in semantics; the semantics of some sentences must be prior to the semantics of other sentences. Only after the sentence "Grass is green" is made true, is there a fact to which the sentence "'Grass is green' is true" can correspond. This I believe to be a lesson of Kripke's grounding approach to semantics. However, as will be explained, the problem with Kripke's theory is that the proposed order is too simplistic. It is the purpose of this paper to propose a modification.

I do not think that a satisfying explanation of what "grounding" consists in, has been given anywhere in the literature. Nor do I think that it is reasonable to take it as a primitive.[2] However, I will leave the task of cashing it out (I think it can be!) for another occasion, and, for the purpose of this paper, assume it to be a meaningful and explanatory metaphysical concept with roughly the properties that its defenders take it to have. In the absence of a literal account of what grounding consists in, I shall require a metaphor. The one just presented suits me well and I shall make extensive use of it.

[2] As argued in [16].

3. Motivation for the modification

The undefinedness of a sentence is a fact that the writer should be able to describe in *The True*. However, he makes the mistake of being far too optimistic about the possibility of sentences becoming either true or false and therefore postponing the assignment of the value of undefined "forever". He could instead make a sentence undefined earlier by writing *The Undefined* in parallel with the two other books and thereby make facts of undefinedness available for sentences to correspond to. He could do that when it becomes clear that there is no longer any hope of the sentence becoming true or false, because there is no progress towards satisfying its truth or falsity conditions. I will explain this idea informally in this section.

From the rules of Kripke's theory, we can "distill" the following informal criteria for making a sentence true, false, and undefined – criteria that will be adopted in this paper. A declarative sentence is true if the claim it expresses is the case "prior" to the sentence getting a truth value. Or to put it a bit differently: a declarative sentence is true if the state of affairs postulated by the sentence to be the case, can obtain in the Kripkean hierachy independently of this sentence itself getting a truth value. Likewise, a declarative sentence is false if the claim it expresses is not-the-case "prior" to the sentence getting a truth value. There is a tertium between these two possibilities, namely that the truth value of the sentence cannot be determined "prior" to this determination itself. In this case, the sentence is undefined.

If we accept this interpretation of the theory and these criteria for the three truth values, Kripke's theory can be criticised for being too optimistic. That is, if the author of great books were to follow Kripke's rules, he would be too optimistic. A principle in Kripke's theory is the following: A sentence can wait arbitrarily long for its truth conditions or falsity conditions to obtain. Call this "the principle of optimism". For a simple example of how the principle works without being problematic, consider the following sentences:

S1: Grass is green
S2: S1 is true
S3: S2 is true

Before the writer has begun work on the books, the greenness of grass is a fact but the truth values of the three sentences are not yet determined. Writing the first chapter of each book, he checks if he can give the sentences truth values. S1 can be written in *The True* as its truth condition already obtains. He cannot give S3 one, for whether or not S2 will occur in *The True* is not yet determined. But according to the principle of optimism, he should just wait and see if he can later. He also cannot give S2 a truth value as

its truth condition did not already obtain (all atomic sentences of a given chapter are to be imagined written simultaneously). After having written the first chapters of the books, S3 still cannot be given a truth value, and so the author just keeps waiting with regard to this sentence. The truth condition of S2 is that S1 is written in *The True*. This condition has obtained prior to the writing of the second chapters, so S2 is included in chapter 2 of *The True*. So finally after having written those chapters, it has been determined in advance that S2 is true, so S3 can be listed in chapter 3 of *The True*.

Following these principles, the Liar is never given a truth value. At every point in the writing process, neither the truth condition (the Liar being included in *The False*) nor the falsehood condition (the Liar being included in *The True*) has previously obtained, so the Liar just waits forever. In Kripke's theory, this is exactly what it takes for a sentence to be declared undefined. Because of this, the theory has its limitations. Consider the sentence

S4: "The Liar is false" is undefined

Intuitively, this sentence is true, but according to Kripke's theory, it is not. The problem is that there is no instant of time after the point when the Liar is given the value "undefined", at which the writer can add S4 as true. We would have to introduce a "meta-writer", i.e. an author who can write about the first author while the latter is somehow banned from writing about the former. This brings us back to the primitive Tarskian approach where the semantic facts about a language can only be stated (and given the right semantics) in another language.

The problem with S4 is that, following Kripke's theory, the writer is too optimistic about the Liar; he keeps hoping forever that it will get a truth value. Using a better theory, he would, at some stage, come to the conclusion that there is no hope for the Liar, so that at some non-ultimate instant of time, the Liar could be given the truth value "undefined" and then at the next instant, S4 could become true, grounded in the fact about this ungroundedness. For this reason, I will replace the principle of optimism with this principle of hope: As long as there is hope that a given sentence can become true or false, we must wait. At such time as there is no longer any hope of that, the sentence is given the truth value "undefined". Of course this principle calls for a clarification: When is there still hope and when is there none?

Consider the example of the sentences S1–S3 again. The reason the writer should not give up on S2 and S3, even though he cannot initially give them a truth value, is that there are sentences, on which S2 and S3 depend, which are getting truth values. S3 depends on S2, which again depends on S1.

So indirectly, S3 also depends on S1. So the reason that the writer should not declare S2 and S3 undefined at level 1 in the hierarchy is that there is a sentence on which they depend, which gets a truth value at that level. Likewise, S3 should not be declared undefined at level 2 because S2 gets a truth value at that level and S3 depends on S2. It is not necessary for the writer to wait until level 3 where S3 actually gets a proper truth value to see that it was reasonable at level 1 and 2 to keep open the possibility of making S3 true or false. That hope is warranted by the fact that at these levels there is progress towards satisfying the truth/falsity conditions of S3. Such hope does not guarantee that a given sentence will become true or false, but the lack of hope at some level is a guarantee that in Kripke's theory it would not.

The Liar, on the other hand, depends only on itself. So at level 1 it cannot be declared true or false, and since there is therefore no sentence on which the Liar depends which does get a truth value at level 1, there is no longer any hope. So at this level, the Liar can be written in *The Undefined* by the writer and then at level 2, he can add S4 to *The True*.

As these examples show, the clarification of the principle of hope presupposes a clarification of an auxiliary notion of dependency. We first define a notion of "direct dependency", and both notions are relative to the level in the iteration. A negation depends directly on the negated sentence if it is "undetermined", i.e. has not yet been given a truth value. A disjunction/conjunction/conditional/bi-conditional depends directly on those of its disjuncts/etc. that are undetermined. A quantified sentence depends directly on those of its instances that are undetermined. And a sentence claiming truth, falsity or undefinedness of some other sentence depends directly on that sentence if it is undetermined. The dependency relation is then simply the transitive closure of the direct dependency relation.

With the concept of dependency, we can formulate this condition (sufficient but not necessary) for maintaining hope that a sentence s will become true or false: As long as more and more of the sentences on which s depends are getting truth values, there is still hope for s.

The example of the Liar and S4 suggests another condition: If there is a sentence s_0 on which s depends such that s_0 does not depend on s, then there is still hope for s. But Yablo's Paradox [21] shows that this condition is too simplistic. For each of the Yablo sentences, there is no hope that it will become true or false even though the suggested condition is satisfied. So this modified condition will be adopted instead: If there is a sentence s_0 on which s depends such that s_0 does not depend on s and there is no infinite sequence $s_0, s_1, s_2, \ldots$ consisting of distinct sentences such that for every $n \in \mathbb{N}_0$, s_n depends on s_{n+1}, then there is still hope for s.

On behalf of the writer, we will adopt one further principle of hope. Consider this pair of sentences (the Indirect Liar):

S5: S6 is false

S6: S5 is true

It would be possible to give one of these sentences a proper truth value (i.e. true or false) by letting that sentence wait longer than the other. But it does not seem reasonable to do so, as both of the two possible choices would be arbitrary. So they should both come out as undefined, and that is also the result with the two principles stated above. But now consider this pair of sentences:

S7: S8 is false

S8: "S8 is false" is true

The structure of the dependence relation is the same as in the former example, and with only the two principles, the result is also the same; both S7 and S8 come out as undefined. However, in this case, the writer can do what he could not in the former, namely let one of the sentences wait longer than the other in a non-arbitrary way. Since S8 quotes S7, it seems quite reasonable to evaluate S7 first. So we adopt this third principle: If s depends on a sentence which does not yet have a truth value and is quoted in s, there is still hope for s. With this principle, S7 becomes undefined and S8 false.

The purpose of replacing the principle of optimism with the first principle of hope is to get some of the sentences which in Kripke's theory would become undefined anyway (in the fixed point) to become undefined earlier. The purpose of adding the second and third principles of hope is to ensure that if a sentence s is about other sentences which are made undefined, and the fact of this undefinedness can exist independently of the truth value of s, then s is allowed to wait for this fact to be created.

It turns out (theorem 5.2) that these three principles of hope taken together are not overtly optimistic. By this I mean that at each level of the hierarchy where there are still sentences which are not either true, false or undefined, some of these must become true or false at that level, or else not satisfy any of the principles of hope and thus be made undefined. And that is enough to ensure that the construction reaches a fixed point where all sentences are either true, false or undefined.

4. Truth values, connectives, quantifiers and quotes

In the formal language defined below, there is a predicate for each of the three truth values: T for truth, F for falsehood and U for undefined.[3] In the meta-language used to specify the semantics, these symbols are used for the truth values: $\top$ for truth, $\bot$ for falsehood and $+$ for undefined. In addition, the symbol $\mid$ is used for undetermined, i.e. for a sentence which at some level has not yet received a truth value. That is, undetermined is not itself a truth value; no sentence will be declared undetermined by the theory presented below. It is a technical device used in the formal construction which, like Kripke's, will result in a trivalent semantics. $\top$, $\bot$, $+$ and $\mid$ are called "semantical values", only $\top$, $\bot$ and $+$ "truth values", and again only $\top$ and $\bot$ "proper truth values". A sentence having the value $+$ means that it was not possible to give it a proper truth value – that it had to be "given up". Having the value $\mid$ means that it has not yet received a truth value – but it will eventually, at a higher level.

We need truth tables for the connectives and semantics for the quantifiers covering all the four semantic values.[4] Let us first consider negation. Of course negation should take $\top$ to $\bot$ and $\bot$ to $\top$. If a sentence is undetermined then so is its negation. So we let negation take $\mid$ to $\mid$. This leaves the case of $+$. Consider the negation of the Liar:

S9: It is not the case that the Liar is false

Since the Liar is undefined, S9 is intuitively true. And since S9 depends on the Liar but not the other way around, that is also how it should come out according to the second "principle of hope" above. So negation should take $+$ to $\top$, i.e. it should be of the exclusion variant. This is also in line with the general condition for truth given in the second paragraph of section 3. According to this condition, a sentence of the form $\neg s$ is true if it is the case prior to $\neg s$ getting a truth value that s is not the case. And if s is undefined, s is not the case.

This gives us the truth table for the negation illustrated in figure 1. This figure also shows the truth tables which will be used for the other connectives. They are all based on the same principles as the one for negation: Begin with

---

[3]The predicate F is needed as a primitive as $F(c)$ is equivalent to neither $\neg T(c)$ nor $T(c')$ for a constant c' denoting the negation of the denotation of c. This is seen from the formal theory presented in the next section, for example by letting c denote the Truth Teller. Then c denotes an undefined sentence, while c' denotes a true one. Hence, $F(c)$ is false and both $\neg T(c)$ and $T(c')$ are true.

[4]Dunn [6] and Belnap [3] have developed semantics for the connectives and quantifiers of a four-valued logic, but as they interpret the fourth value as "both true and false" we cannot lean on them here.

$\neg\phi$	
$\top$	$\bot$
$\bot$	$\top$
$+$	$\top$
$\mid$	$\mid$

$\phi \vee \psi$	$\top$	$\bot$	$+$	$\mid$
$\top$	$\top$	$\top$	$\top$	$\top$
$\bot$	$\top$	$\bot$	$\bot$	$\mid$
$+$	$\top$	$\bot$	$\bot$	$\mid$
$\mid$	$\top$	$\mid$	$\mid$	$\mid$

$\phi \wedge \psi$	$\top$	$\bot$	$+$	$\mid$
$\top$	$\top$	$\bot$	$\bot$	$\mid$
$\bot$	$\bot$	$\bot$	$\bot$	$\bot$
$+$	$\bot$	$\bot$	$\bot$	$\bot$
$\mid$	$\mid$	$\bot$	$\bot$	$\mid$

$\phi \rightarrow \psi$	$\top$	$\bot$	$+$	$\mid$
$\top$	$\top$	$\bot$	$\bot$	$\mid$
$\bot$	$\top$	$\top$	$\top$	$\top$
$+$	$\top$	$\top$	$\top$	$\top$
$\mid$	$\top$	$\mid$	$\mid$	$\mid$

$\phi \leftrightarrow \psi$	$\top$	$\bot$	$+$	$\mid$
$\top$	$\top$	$\bot$	$\bot$	$\mid$
$\bot$	$\bot$	$\top$	$\top$	$\mid$
$+$	$\bot$	$\top$	$\top$	$\mid$
$\mid$	$\mid$	$\mid$	$\mid$	$\mid$

Figure 1.

the classical truth tables and treat $+$ as $\bot$. When there is enough "information in the input" to determine the truth value to be $\top$ and $\bot$ respectively, then $\top$ and $\bot$ respectively is the output. If there is not enough information, the output is $\mid$.[5]

This has the consequence that $\phi \rightarrow \psi$ means "if ϕ is true then ψ is as well", securing the validity of modus ponens, and that $\phi \leftrightarrow \psi$ means "(if ϕ is true then ψ is as well) and (if ψ is true then ϕ is as well)" and not that ψ and ϕ are equivalent in the sense of having the same truth value.

The semantics of the existential and universal quantifiers will – as in classical logic – be determined by treating them as infinite disjunction and infinite conjunction respectively.

As motivated by examples like S7 and S8, the language will be equipped with a device for quotation. S7 can be formalised as $F(c_8)$ where c_8 is a constant referring to this sentence, the formalization of S8: $T(\ulcorner F(c_8) \urcorner)$. That is, the quoted sentence appears as a concrete syntactical part of the sentence.[6]

[5] The truth tables for conjunction, implication and bi-implication are as they would be if these connectives where defined in the usual way from negation and disjunction. However, they will not be defined like this because this would increase the number of sentences that a complex sentence depends on. For instance, a sentence of the form $\phi \rightarrow \psi$ would depend on $\neg\phi$. As a consequence, theorem 9.2 below would fail.

[6] Note that this means that the symbols "$\ulcorner$" and "$\urcorner$" are in the object language and are not used as a meta-language codification device as elsewhere in the literature.

5. The formal theory

For each $n \in \mathbb{N}$, let there be a set $\mathcal{P}_n$ of *ordinary n-ary predicates*. In addition, we have the *extra-ordinary predicates*: T (true), F (false) and U (undefined). We also have a set $\mathcal{C}$ of *constants* and a set of *variables*.

The set of *well-formed formulas (wff's)* and the set of *terms* are defined by simultaneous recursion, like this:

- If c is a constant, v a variable and ϕ a wff, then c, v and $\ulcorner\phi\urcorner$ are terms.
- If P is an ordinary n-ary predicate and $t_1, \ldots, t_n$ are terms, then $P(t_1, \ldots, t_n)$ is a wff.
- If ϕ and ψ are wff's, then $\neg\phi$, $(\phi \vee \psi)$, $(\phi \wedge \psi)$, $(\phi \rightarrow \psi)$ and $(\phi \leftrightarrow \psi)$ are wff's.
- If ϕ is a wff and v a variable, then $\exists v\phi$ and $\forall v\phi$ are wff's.
- If t is a term, then $T(t)$, $F(t)$ and $U(t)$ are wff's.
- Nothing is a wff or a term except by virtue of the above clauses.

Brackets will be suppressed when no confusion may arise.

Given a wff ϕ, a variable v and a constant c, we understand by $\phi(v/c)$ the wff which is identical with ϕ with the possible exception that all free occurrences of v are replaced with c ("free occurrences of variables" is defined as is usual; in particular, there is no difference between what counts as a free occurrence of a variable in ϕ and $\ulcorner\phi\urcorner$).

For wff's ϕ and ϕ', we say that ϕ' *is quoted in* ϕ if there is a wff ϕ'' such that $\ulcorner\phi''\urcorner$ appears in ϕ, and ϕ' is the result of replacing free occurrences of variables in ϕ'' with constants, i.e. if ϕ' is of the form $\phi''(v_1/c_1)\cdots(v_n/c_n)$, where $n \in \mathbb{N}_0$, $v_1, \ldots, v_n$ are variables and $c_1, \ldots, c_n$ are constants.[7]

A wff is a *sentence* and a term is called *closed* if they do not contain any free occurrences of variables. Let $\mathcal{S}$ be the set of sentences and $\mathcal{CT}$ the set of closed terms. That concludes the specification of the syntax of the language.

A *model* is a pair $\mathfrak{M} = (D, I)$ such that

- D, the *domain*, is a superset of $\mathcal{S}$, and
- I, the *interpretation function*, is a function defined on $\bigcup_{n\in\mathbb{N}} \mathcal{P}_n \cup \mathcal{CT}$ such that
  - for every $P \in \mathcal{P}_n$, $I(P) \subseteq D^n$,
  - for every $c \in \mathcal{C}$, $I(c) \in D$,

[7] Note that in the case $n = 0$, i.e. when $\phi' = \phi''$, we have what would normally be called a quote; ϕ' appears as a concrete syntactical part of ϕ inside quotation marks. The cases where $n > 0$ loosens this a bit: An infinity of sentences can be quoted at once if they only differ in what constants they contain. Yet, it is not loosened so much that "quoted in" means anything like "refers to" which is a much more inclusive relation. So while self-reference is possible, self-quotation is (naturally enough) not.

– $I[\mathcal{C}] = D$, and
– for every $s \in \mathcal{S}$, $I(\ulcorner s \urcorner) = s$.[8]

Let a model be fixed for the remainder of this paper. In order to avoid having to get into the technicalities of arithmetization and diagonalization, we will simply make certain assumptions about the model when needed, such as that there is a sentence $F(c_l)$ where c_l is a constant such that $I(c_l) = F(c_l)$ to have a liar sentence.

We call a triple $\mathcal{E} = (\mathcal{T}, \mathcal{F}, \mathcal{U})$ such that $\mathcal{T}$, $\mathcal{F}$ and $\mathcal{U}$ are subsets of $\mathcal{S}$ an *evaluation*.We say that $\mathcal{E}$ is *consistent* if $\mathcal{T}$, $\mathcal{F}$ and $\mathcal{U}$ are disjoint, and *total* if $\mathcal{T} \cup \mathcal{F} \cup \mathcal{U} = \mathcal{S}$. We also say that an evaluation $\mathcal{E}' = (\mathcal{T}', \mathcal{F}', \mathcal{U}')$ *extends* $\mathcal{E}$ if $\mathcal{T} \subseteq \mathcal{T}'$, $\mathcal{F} \subseteq \mathcal{F}'$ and $\mathcal{U} \subseteq \mathcal{U}'$. For each set $S \subseteq \mathcal{S}$ of sentences, define $\mathcal{E}|S := (\mathcal{T} \cap S, \mathcal{F} \cap S, \mathcal{U} \cap S)$.

Before continuing with the definitions, here is an informal explanation of the idea behind them. The semantics is built up through levels like in Kripke's theory. At the "beginning of a level", there are (possibly) some sentences that have already received a truth value, i.e. there is an initial evaluation in the sense just defined, in which the first set in the triple is the set of true sentences, the second is the set of false sentences and the third is the set of undefined sentences. At the "end of a level", more sentences have received truth values (unless the initial evaluation is total). A level consists of two parts. In the first, new true and false sentences are added. The result is called a "tentative evaluation". In the second, new sentences are declared undefined and the result is called an "evaluation".

We define the *tentative evaluation with respect to the evaluation* $\mathcal{E} = (\mathcal{T}, \mathcal{F}, \mathcal{U})$, $\mathrm{TE}_{\mathcal{E}}$, as $(\mathcal{T}_{\mathcal{E}}, \mathcal{F}_{\mathcal{E}}, \mathcal{U})$, where $\mathcal{T}_{\mathcal{E}} = \mathcal{T} \cup t_{\mathcal{E}}$ and $\mathcal{F}_{\mathcal{E}} = F \cup f_{\mathcal{E}}$, where again $t_{\mathcal{E}}$ and $f_{\mathcal{E}}$ are defined by recursion on the complexity[9] of the sentence as follows:

TE1) If s is of the form $P(t_1, \ldots, t_n)$ where P is an ordinary n-ary predicate and $t_1, \ldots, t_n$ are closed terms, then
- $s \in t_{\mathcal{E}}$ if $(I(t_1), \ldots, I(t_n)) \in I(P)$, and
- $s \in f_{\mathcal{E}}$ otherwise.

TE2) If s is of the form $\neg\phi$ where ϕ is a sentence and $s \notin \mathcal{U}$, then

[8] The quotation device provides for an intensional context, so that for a suitably defined binary identity predicate, a sentence of the form $=(c, c') \wedge \neg =(\ulcorner P(c) \urcorner, \ulcorner P(c') \urcorner)$ can be true, reflecting the natural language truth "Cicero is identical to Tully, but 'Cicero is a great orator' is different from 'Tully is a great orator'".

[9] In a sense of "complexity" where any sentence in which a predicate (ordinary or extra-ordinary) has widest scope is of minimal complexity, while sentences where a connective or quantifier has widest scope is of higher complexity than what is in the scope. That is, a sentence which is about another sentence being, say, true, is of minimal complexity even if it quotes that other sentence and that sentence is of high complexity.

- $s \in t_{\mathcal{E}}$ if $\phi \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$, and
- $s \in f_{\mathcal{E}}$ if $\phi \in \mathcal{T}_{\mathcal{E}}$.

TE3) If s is of the form $(\phi \vee \psi)$ where ϕ and ψ are sentences and $s \notin \mathcal{U}$, then
- $s \in t_{\mathcal{E}}$ if $\phi \in \mathcal{T}_{\mathcal{E}}$ or $\psi \in \mathcal{T}_{\mathcal{E}}$, and
- $s \in f_{\mathcal{E}}$ if $\phi \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$ and $\psi \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$.

TE4) [Analogous for $(\phi \wedge \psi)$, see section 4]

TE5) [Analogous for $(\phi \rightarrow \psi)$]

TE6) [Analogous for $(\phi \leftrightarrow \psi)$]

TE7) If s is of the form $\exists v\phi$ where v is a variable and ϕ is a wff with at most v free and $s \notin \mathcal{U}$, then
- $s \in t_{\mathcal{E}}$ if there exists a $c \in \mathcal{C}$ such that $\phi(v/c) \in \mathcal{T}_{\mathcal{E}}$, and
- $s \in f_{\mathcal{E}}$ if for all $c \in \mathcal{C}$, $\phi(v/c) \in \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$.[10]

TE8) [Analogous for $\forall v\phi$]

TE9) If s is of the form $T(t)$ where t is a closed term and $s \notin \mathcal{U}$, then
- $s \in t_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{T}$,
- $s \in f_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{F} \cup \mathcal{U}$, and
- $s \in f_{\mathcal{E}}$ if there is no sentence s' such that $I(t) = s'$.

TE10) If s is of the form $F(t)$ where t is a closed term and $s \notin \mathcal{U}$, then
- $s \in t_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{F}$,
- $s \in f_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{T} \cup \mathcal{U}$, and
- $s \in f_{\mathcal{E}}$ if there is no sentence s' such that $I(t) = s'$.

TE11) If s is of the form $U(t)$ where t is a closed term and $s \notin \mathcal{U}$, then
- $s \in t_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{U}$,
- $s \in f_{\mathcal{E}}$ if there is a sentence s' such that $I(t) = s'$ and $s' \in \mathcal{T} \cup \mathcal{F}$, and
- $s \in f_{\mathcal{E}}$ if there is no sentence s' such that $I(t) = s'$.

Next, we formalise the notion of dependency discussed above. We make the relation relative to evaluations, for as more and more sentences get truth values, there are fewer and fewer that can still affect the truth value of a given sentence. The binary relation $R_{\mathcal{E}}$ on $\mathcal{S}$, called the *direct dependency relation with respect to the evaluation* $\mathcal{E}$, is defined as follows: $sR_{\mathcal{E}}s'$ if both s and s' are undetermined according to $\mathcal{E}$, and s is $\neg\phi$ and s' is ϕ, or s is $\phi \vee \psi$, $\phi \wedge \psi$, $\phi \rightarrow \psi$ or $\phi \leftrightarrow \psi$ and s' is ϕ or ψ, or s is $\exists v\phi$ or $\forall v\phi$ and s' is

[10] This has the consequence that the truth value of $\exists v\phi$ depends on the truth values of the sentences of the form $\phi(v/c)$ as they have been assigned at various earlier levels, rather than on whether the objects of the domain satisfy ϕ at the level where $\exists v\phi$ is assigned a truth value. There is room for variation here, but I will not go into the alternatives.

$\phi(v/c)$ for some constant c, or s is $T(t)$, $F(t)$ or $U(t)$ and s' is $I(t)$.

Let $\overline{R}_{\mathcal{E}}$, the *dependency relation with respect to the evaluation* $\mathcal{E}$, be the transitive closure of $R_{\mathcal{E}}$. For $s \in \mathcal{S}$, define $\overline{R}_{\mathcal{E}}(s) := \{s' | s \overline{R}_{\mathcal{E}} s'\}$.[11]

We can now define "evaluation". We want to add sentences for which there is no longer any hope to the set of undefined sentences. The conditions E2, E3 and E4 below correspond directly to the three principles of hope introduced in section 3, in the order they were mentioned.

Set the *evaluation with respect to the evaluation* $\mathcal{E} = (\mathcal{T}, \mathcal{F}, \mathcal{U})$, $\mathrm{E}_{\mathcal{E}}$, equal to $(\mathcal{T}_{\mathcal{E}}, \mathcal{F}_{\mathcal{E}}, \mathcal{U}_{\mathcal{E}})$, where $\mathcal{U}_{\mathcal{E}}$ is the union of $\mathcal{U}$ and the set of sentences s such that

E1) $s \notin \mathcal{T}_{\mathcal{E}} \cup \mathcal{F}_{\mathcal{E}} \cup \mathcal{U}$,
E2) $\mathrm{TE}_{\mathcal{E}} | \overline{R}_{\mathcal{E}}(s) = \mathcal{E} | \overline{R}_{\mathcal{E}}(s)$,
E3) for every sentence s_0, if $s \overline{R}_{\mathcal{E}} s_0$ then ($s_0 \overline{R}_{\mathcal{E}} s$ or there is an infinite $\overline{R}_{\mathcal{E}}$-sequence $s_0 \overline{R}_{\mathcal{E}} s_1 \overline{R}_{\mathcal{E}} s_2 \overline{R}_{\mathcal{E}} \cdots$ consisting of distinct elements), and
E4) there is no sentence s' which is quoted in s such that $s \overline{R}_{\mathcal{E}} s'$.

To get the final evaluation, we simply iterate the process of making interpretations, beginning with the empty evaluation and continuing until every sentence has a truth value. For all ordinals α, define the *evaluation with respect to the level* α, E^{α}, by recursion:

$$\mathrm{E}^{\alpha} = \begin{cases} (\emptyset, \emptyset, \emptyset) & \text{if } \alpha = 0 \\ \mathrm{E}_{\mathrm{E}^{\alpha-1}} & \text{if } \alpha \text{ is a succesor ordinal} \\ \left(\bigcup_{\eta<\alpha} \mathcal{T}_{\mathrm{E}^{\eta}}, \bigcup_{\eta<\alpha} \mathcal{F}_{\mathrm{E}^{\eta}}, \bigcup_{\eta<\alpha} \mathcal{U}_{\mathrm{E}^{\eta}}\right) & \text{if } \alpha \text{ is a limit ordinal} \neq 0 \end{cases}$$

For each ordinal α, let $\mathcal{T}^{\alpha}$, $\mathcal{F}^{\alpha}$ and $\mathcal{U}^{\alpha}$ be the sets such that E^{α} equals $(\mathcal{T}^{\alpha}, \mathcal{F}^{\alpha}, \mathcal{U}^{\alpha})$.

No inconsistencies arise in the iterative process. That is the content of this lemma:

LEMMA 5.1. *For every model, all of the evaluations* $\mathrm{TE}_{\mathrm{E}^{\alpha-1}}$ *for every successor ordinal* α *and* E^{α} *for every ordinal* α *are consistent.*

PROOF. This will be proved by induction on the sequence of evaluations, including the tentative, in the order in which they appear in the above construction. First note that this sequence is monotone in the sense that each of the three sets in a given evaluation in the sequence is a superset of the corresponding set in any evaluation earlier in the sequence. This fact, which is immediate from the definitions of tentative evaluation and evaluation, will be

[11] For other definitions of dependency see [4], [14] and [20]. In these papers, however, dependency is only used to *analyze* truth and paradox and to delimit "safe" fragments of languages that allow for self-reference, not to influence the assignment of truth values.

used twice in this proof and multiple times thereafter, often without explicit mention.

The base case is trivial as $\mathrm{E}^0 = (\emptyset, \emptyset, \emptyset)$.

In the induction step, first take the case of $\mathrm{TE}_{\mathrm{E}^{\alpha-1}} = (\mathcal{T}^\alpha, \mathcal{F}^\alpha, \mathcal{U}^{\alpha-1})$ where α is a successor ordinal. Consider the set S of sentences which are in more than one of the sets $\mathcal{T}^\alpha$, $\mathcal{F}^\alpha$ and $\mathcal{U}^{\alpha-1}$. We must prove that S is empty. By the induction hypothesis, none of the sentences in S are in more than one of the sets $\mathcal{T}^{\alpha-1}$, $\mathcal{F}^{\alpha-1}$ and $\mathcal{U}^{\alpha-1}$. From TE1 it is seen that none of the elements of S can be of the form $P(t_1, \ldots, t_n)$ where P is an ordinary n-ary predicate. All the other clauses (TE2–TE11) contain the assumption that the relevant sentence is not in $\mathcal{U}^{\alpha-1}$. Ergo, S must consist entirely of sentences which are in both $\mathcal{T}^\alpha$ and $\mathcal{F}^\alpha$. However, this implies that for each $s \in S$ both the condition for s being in $t_{\mathrm{E}^{\alpha-1}}$ and the condition for s being in $f_{\mathrm{E}^{\alpha-1}}$ by the relevant one of TE2–TE11 are satisfied – since s being in $\mathcal{T}^{\alpha-1}$ implies the condition for s being in $t_{\mathrm{E}^{\alpha-1}}$ being satisfied, by the monotonicity of the truth, falsehood and undefined sets, and likewise for s being in $\mathcal{F}^{\alpha-1}$. Upon inspection of TE2–TE11, it is seen that this implies that s being in both $\mathcal{T}^\alpha$ and $\mathcal{F}^\alpha$ is conditioned upon some other sentence being in both $\mathcal{T}^\alpha$ and $\mathcal{F}^\alpha$ (TE2–TE8) or in both $\mathcal{T}^{\alpha-1}$ and $\mathcal{F}^{\alpha-1}$ (TE9–TE11). And then the recursive nature of the definition of $t_{\mathrm{E}^{\alpha-1}}$ and $f_{\mathrm{E}^{\alpha-1}}$ implies that S is empty.

Next, take the case of the E^α's for α a successor ordinal. These are the evaluations that are "based on" a tentative evaluation, and the induction hypothesis is that this tentative evaluation is consistent. So this step is easy: From the definition of evaluation with respect to a model and an evaluation, it is seen that only sentences that are in neither the "truth set" nor the "falsehood set" are added to the "undefined set", so no inconsistency is created.

The last case is the E^α's for α a limit ordinal different from 0. E^α is equal to $\left(\bigcup_{\eta<\alpha} \mathcal{T}_{\mathrm{E}^\eta}, \bigcup_{\eta<\alpha} \mathcal{F}_{\mathrm{E}^\eta}, \bigcup_{\eta<\alpha} \mathcal{U}_{\mathrm{E}^\eta}\right)$. Assume *ad absurdum* that this is inconsistent. Then there is a sentence s which is in both, say, $\bigcup_{\eta<\alpha} \mathcal{T}_{\mathrm{E}^\eta}$ and $\bigcup_{\eta<\alpha} \mathcal{F}_{\mathrm{E}^\eta}$ (the other possibilities are of course analogous). It follows that there are η' and η'' smaller than α such that $s \in \mathcal{T}_{\mathrm{E}^{\eta'}}$ and $s \in \mathcal{F}_{\mathrm{E}^{\eta''}}$. Let η''' be equal to the largest of η' and η''. By monotonicity, it follows that $s \in \mathcal{T}_{\mathrm{E}^{\eta'''}}$ and $s \in \mathcal{F}_{\mathrm{E}^{\eta'''}}$, which contradicts the induction hypothesis. This concludes the proof. □

For all sentences s and ordinals α we define the *interpretation of s with respect to the level α*, written $[\![s]\!]^\alpha$, as follows:

$$[\![s]\!]^\alpha = \begin{cases} \top & \text{if } s \in \mathcal{T}^\alpha \\ \bot & \text{if } s \in \mathcal{F}^\alpha \\ + & \text{if } s \in \mathcal{U}^\alpha \\ | & \text{if } s \notin \mathcal{T}^\alpha \cup \mathcal{F}^\alpha \cup \mathcal{U}^\alpha \end{cases}$$

The following theorem shows that the process will actually result in all the sentences getting a truth value.

THEOREM 5.2. *For every model, a unique total evaluation $\mathcal{E}$ exists such that for some ordinal α, $\mathrm{E}^\alpha = \mathcal{E}$. At all higher levels, the evaluation is the same, i.e. for all ordinals $\beta > \alpha$, $\mathrm{E}^\beta = \mathcal{E}$.*

PROOF. For any evaluation $\mathcal{E}'$, $\mathrm{E}_{\mathcal{E}'}$ is an extension of $\mathcal{E}'$. So the second claim of the theorem follows from the first together with lemma 5.1. And this observation also implies the uniqueness part of the first claim.

We will prove that for every ordinal α, if E^α is not total then $\mathrm{E}^{\alpha+1} \neq \mathrm{E}^\alpha$. As the sentences form a set, this implies the existence part of the first claim.

Assume toward a contradiction that there is an α such that E^α is not total but $\mathrm{E}^{\alpha+1} = \mathrm{E}^\alpha$. Let s be a sentence that is not in any of the three sets in E^α. The contradiction will be established by finding a sentence that satisfies all of the conditions E1–E4. Such a sentence will be found by using what can very informally be described as dependency chains which start with s, are infinitely long if possible, reluctant to repeat themselves, and if it has to repeat itself it will do so by using an element whose copies are as far back in the sequence as possible. More precisely, we will consider sequences $\langle s_\gamma | \gamma < \eta \rangle$ with the following characteristics:

a) $1 \leq \eta \leq \omega$.
b) $s_0 = s$.
c) Every s_γ is a sentence.
d) For every γ such that $\gamma + 1 < \eta$, $s_\gamma \overline{R}_{\mathrm{E}^\alpha} s_{\gamma+1}$ holds. In addition, for each $\beta \leq \gamma$, if $s_{\gamma+1} = s_\beta$, then for all s' such that $s_\gamma \overline{R}_{\mathrm{E}^\alpha} s'$ it is the case that $s' = s_\delta$ for some $\beta \leq \delta \leq \gamma$.
e) $\eta < \omega$ only if there is no sentence s' such that $s_{\eta-1} \overline{R}_{\mathrm{E}^\alpha} s'$.

Let us call such sequences *relevant* for want of a better word. We first prove that at least one exists by constructing one by recursion. The base case is obvious: $s_0 = s$. Then for a given finite ordinal γ, assume that s_γ has been chosen. If there is no sentence s' such that $s_\gamma \overline{R}_{\mathrm{E}^\alpha} s'$, then by clause e) the construction is completed. If there is, we need to pick one of them as $s_{\gamma+1}$. If there is one that does not appear earlier in the sequence, pick one of those. If not, pick the s' for which $\max\{\delta | s_\delta = s'\}$ is minimal. Then all the clauses are satisfied.

By the definition of the direct dependency relation, only sentences s' such that $s' \notin \mathcal{T}^{\alpha+1} \cup \mathcal{F}^{\alpha+1} \cup \mathcal{U}^{\alpha}$ can be the second relata of R_{E^α}. So it is also just these sentences that can be the second relata of $\overline{R}_{\mathrm{E}^\alpha}$. Ergo, for every element s' of a relevant sequence $s' \notin \mathcal{T}^{\alpha+1} \cup \mathcal{F}^{\alpha+1} \cup \mathcal{U}^{\alpha} = \mathcal{T}_{\mathrm{E}^\alpha} \cup \mathcal{F}_{\mathrm{E}^\alpha} \cup \mathcal{U}^{\alpha}$ holds, i.e. E1 is satisfied. But we also have $\mathrm{TE}_{\mathrm{E}^\alpha} = \mathrm{E}^\alpha$ and hence $\mathrm{TE}_{\mathrm{E}^\alpha}|\overline{R}_{\mathrm{E}^\alpha}(s') = \mathrm{E}^\alpha|\overline{R}_{\mathrm{E}^\alpha}(s')$, so E2 is also satisfied. Hence, we just need to find an element of a relevant sequence that satisfies E3 and E4. At least one of these statements is true:

1) Every relevant sequence is infinite and contains infinitely many different elements, and there is such a sequence.
2) There is a relevant sequence which is infinite but only contains finitely many different elements.
3) There is a finite relevant sequence.

First assume 1). In one of the sequences there must be an element which satisfies E4; for otherwise an infinite sequence would exist, every element of which, except the first, is quoted in the previous element, i.e. an infinite sequence of shorter and shorter sentences (here "shorter" is not to be interpreted in terms of the concept of complexity described in footnote 9, but simply as "consisting of fewer primitive symbols"). Let s_δ be such an element. s_δ also satisfies E3. For assume that s' is a sentence such that $s_\delta \overline{R}_{\mathrm{E}^\alpha} s'$. Then it follows from 1) and the transitivity of the dependence relation that there is an infinite $\overline{R}_{\mathrm{E}^\alpha}$-sequence $s' \overline{R}_{\mathrm{E}^\alpha} s'_1 \overline{R}_{\mathrm{E}^\alpha} s'_2 \overline{R}_{\mathrm{E}^\alpha} \cdots$ consisting of distinct elements. This concludes the treatment of case 1).

Now assume 2) and let $\langle s_\gamma | \gamma < \omega \rangle$ be a relevant sequence which is infinite but only contains finitely many different elements. Of the elements of $\langle s_\gamma | \gamma < \omega \rangle$, there must be some which are repeated infinitely often. Let s_δ be the shortest of these (or one of them if there are more than one of minimal length). E3 is satisfied for s_δ, for if there were a sentence s' such that $s_\delta \overline{R}_{\mathrm{E}^\alpha} s'$ and not $s' \overline{R}_{\mathrm{E}^\alpha} s_\delta$, then s' would appear in the sequence after some instance of s_δ (by assumption 2) and clause d)), and then s_δ would not appear again, which is a contradiction. E4 is also satisfied for s_δ, for if there were a sentence s' such that s' is quoted in s_δ and $s_\delta \overline{R}_{\mathrm{E}^\alpha} s'$, then s' would be repeated infinitely often in $\langle s_\gamma | \gamma < \omega \rangle$, contradicting the assumption that s_δ is of minimal length among the infinitely often repeated elements. Thus case 2) has been dealt with.

Finally, assume 3) and let $\langle s_\gamma | \gamma < \eta \rangle$ be a finite relevant sequence. It follows directly from clause e) that E3 and E4 are satisfied for the final element of this sequence, $s_{\eta-1}$. This concludes the treatment of case 3) and hence the proof. □

This theorem implies that the theory does not introduce a fourth truth

value. So the inexpressibility of facts about the undefinedness of sentences, that mars Kripke's theory, is not replaced by inexpressibility of facts about a new value.

Letting $\mathcal{E}$ and α be as in the theorem, we can define the *evaluation*, E, as $\mathcal{E}$, and for all sentences s set $[\![s]\!]$ equal to $[\![s]\!]^\alpha$. The value of $[\![s]\!]$ is of course to be thought of as *the* truth value of s.

6. Expressibility of all individual semantic facts

The formalization of the (simple) Liar is $F(c_l)$, where c_l is a constant such that $I(c_l) = F(c_l)$ (i.e. assume that the model satisfies this). At level 1, neither the truth condition nor the falsity conditions of TE10 are satisfied, so $F(c_l)$ is not given a truth value by the tentative evaluation at level 1. $F(c_l)$ only depends on itself (i.e. $\overline{R}_{(\emptyset,\emptyset,\emptyset)}(F(c_l)) = \{F(c_l)\}$), so all of the conditions E1–E4 are satisfied. Ergo, $[\![F(c_l)]\!] = [\![F(c_l)]\!]^1 = +$. The fact that it is undefined can be expressed in the object-language with the sentence $U(c_l)$, which is not given a truth value at level 1 (TE11 and E3) and is made true at level 2.

In order to demonstrate how quantification into a quote can be used to express the undefinedness of infinitely many sentences at a time, and simply because it is a good expository example, let us have a look at how the theory deals with Yablo's Paradox [21]. The paradox results from this infinite list of sentences:

Y1: For all $n > 1$ the sentence Yn is not true
Y2: For all $n > 2$ the sentence Yn is not true
Y3: For all $n > 3$ the sentence Yn is not true
$\vdots$

For each $n \in \mathbb{N}$ the formalisation of Yn is

$$\forall x(P(\bar{n}, x) \to \neg T(x)), \qquad (\mathrm{Y}n)$$

where $\bar{n}$ is a numeral for n, i.e. $\bar{n}$ is a constant such that $I(\bar{n}) = n$, and P is a binary predicate such that

$$I(P) = \{(n, (\mathrm{Y}m)) \mid n, m \in \mathbb{N} \text{ and } m > n\}.$$

At level 1, none of these formulas become true or false (TE1, TE2, TE8, and TE9). At level 1, (Y1) depends on for example the sentence $P(\bar{3}, c_2) \to \neg T(c_2)$, where c_2 is a constant such that $I(c_2) = (\mathrm{Y}2)$. This sentence becomes true at level 1 as the antecedent is false. There are similar sentences for the other Yablo sentences. So E2 is not satisfied for any of the (Yn)'s, and they do not become undefined at level 1 either. At level 2, they also don't get

a proper truth value. At this level, the dependency structure for the Yablo sentences is as illustrated in figure 2. From this figure, it can be seen that all of the clauses E1–E4 are satisfied since it is possible for each (Yn) to construct an infinite and non-repetitive $\overline{R}_{E^1}$-sequence beginning with (Yn). Hence at level 2, they all get the truth value $+$.

The undefinedness of the sentences of Yablo's Paradox can be expressed with the sentence

$$\forall y(N(y) \rightarrow U(\ulcorner \forall x(P(y,x) \rightarrow \neg T(x)))\urcorner),$$

where N is a unary predicate meaning "is a natural number", i.e. $I(N) = \mathbb{N}$. It does not get a truth value at level 1 or 2 (TE1, TE5, TE8, TE11, and E4). At level 3, it gets the value $\top$.

7. Revenge

I promised that the theory would offer expressibility of *all* individual semantic facts, and the simple Liar and Yablo's Paradox were mere examples. I need to show that the promise has been made good on. However, doing so is best coupled with a discussion of the revenge problem. Let us do so informally first:

The Strengthened Liar: The Strengthened Liar is not true

None of the principles of hope applies to the Strengthened Liar, so it is made undefined by the writer of the three books. It is normally considered a failure of a truth value gap theory when this sentence comes out as undefined. For being undefined is to be not true, so the Strengthened Liar being undefined seems to imply that the Strengthened Liar is true. However, that is not the case when the truth criterion is the one introduced in the second paragraph of section 3, namely that a sentence is true if its truth maker can exist independently of and prior to the sentence itself getting a truth value. The potential truth maker for the Strengthened Liar would be that the Strengthened Liar is false or undefined and such a fact cannot exist independently of and prior to the Strengthened Liar getting a truth value. The Strengthened Liar cannot correspond to or represent a fact given independently of its own semantics. Therefore, it is quite reasonable to have it come out as undefined. However, the sentence

S10: "The Strengthened Liar is not true" is not true

which expresses the same claim as the Strengthened Liar, waits for the Strengthened Liar to become undefined before getting a truth value according to the last principle of hope, and hence it becomes true. Strange as it

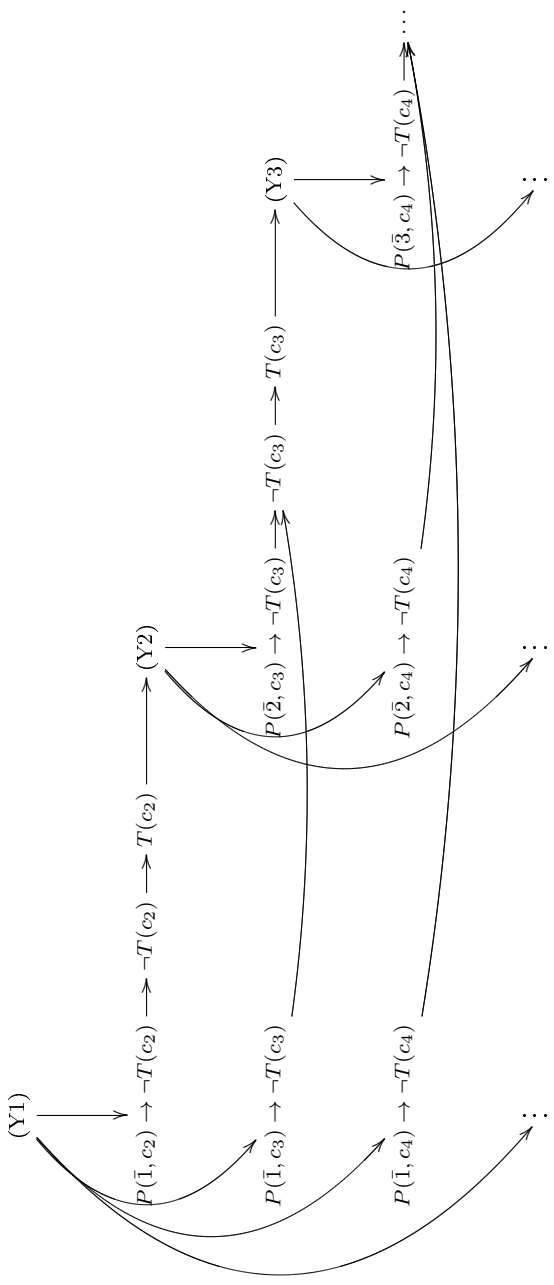

Figure 2. The figure shows the "initial part" of $\overline{R}_{E^1}((\mathrm{Y1}))$, i.e. the dependency structure of the Yablo sentences at level 2. For each $n \in \mathbb{N}$, c_n is a constant such that $I(c_n) = (\mathrm{Y}n)$. The dependency structure is more complicated than shown in this figure if there are more than one constant denoting the same Yablo sentence, but the difference is of no importance.

may seem, this is also quite reasonable; the common claim of the Strengthened Liar and S10 *is* the case at a lower level of the hierarchy than the level where S10 gets a truth value. The potential truth maker for S10 is the same as the potential truth maker for the Strengthened Liar, but it can exist independently of and prior to S10 getting a truth value.

The formalization of the Strengthened Liar is $\neg T(c_s)$, where c_s is a constant such that $I(c_s) = \neg T(c_s)$. Neither $\neg T(c_s)$ nor $T(c_s)$ becomes true or false at level 1 (TE2 and TE9). We have $\overline{R}_{(\emptyset,\emptyset,\emptyset)}(\neg T(c_s)) = \overline{R}_{(\emptyset,\emptyset,\emptyset)}(T(c_s)) = \{\neg T(c_s), T(c_s)\}$, so E1–E4 are satisfied. Ergo, $[\![\neg T(c_s)]\!] = [\![\neg T(c_s)]\!]^1 = +$. If there is some other constant c_r such that $I(c_r) = \neg T(c_s)$ then this fact can be expressed with the sentence $\neg T(c_r)$: It does not get a truth value at level 1 (TE2, TE9 and E3), but at level 2, it gets the value $\top$ (TE2 and TE9). And if nothing else, the formalization of S10, $\neg T(\ulcorner \neg T(c_s) \urcorner)$, can be used to express the fact. Likewise it is left undecided at level 1 (TE2, TE9 and E3/E4), and is made true at level 2 (TE2 and TE9).

This corresponds to the following when we translate back to natural language: "The Strengthened Liar is not true" is named "the Strengthened Liar" and could also have another name, e.g. "the Revenge Liar". "The Strengthened Liar is not true" is undefined but "The Revenge Liar is not true" is true. And if there are no other means to express the fact of this undefinedness, then the sentence "'The Strengthened Liar is not true' is not true" will do the job. Similar use of quotation will always work to express individual semantic facts, since quotation names always exist in natural language and self-quotation is impossible.

So in this theory, unlike in Kripke's, wherein the problem is prevented from arising by expressive weakness, the Strengthened Liar can be expressed but inconsistency is avoided through a certain interpretation of the truth value predicates. The Strengthened Liar is undefined because what it means for a sentence to be undefined is that the truth value of the sentence cannot be determined prior to this determination itself. It is not *also* true because what it means for a sentence to be true is that there is some fact in which the truth can be grounded and which therefore must exist independently of that sentence's having a truth value, and there is not in the case of this sentence. And the undefinedness of the Strengthened Liar can be expressed in the language itself because there are levels after the Strengthened Liar has received its truth value at which the undefinedness of that sentence is a fact to which other sentences can correspond. And the device of quotes ensures that there is at least one suitable, not-yet-determined sentence available at that level.

8. Intersubstitutivity

The extra expressive strength of this theory, compared to Kripke's, comes at the expense of intersubstitutivity of co-referential terms *salva veritate*, a principle to which the pair of sentences $\neg T(c_s)$ and $\neg T(c_r)$ is a counterexample. There is a certain similarity between this theory and contextualism about truth[12], according to which two tokens of the same sentence can have different truth values if they are evaluated in different contexts. In the story of the writer of the great books there are also different contexts, namely the context where noting has been written, the context where only the first chapters have been written, the context where only the first two chapters of each book have been written, and so on. These contexts are more and more inclusive. So the difference between, on the one hand, Kripke's theory and the present, and on the other, contextualism about truth is that a sentence waits for a "favorable" context in which it can get a proper truth value, and if there is such a context it maintains that truth value in all later contexts.

This context dependence opens up for the possibility that two sentences which are identical except for substitution of co-referential terms can have different truth values if they are evaluated in different contexts. If substitution of a term in a self-referential sentence with another with the same reference results in a non-self-referential sentence, the two sentences may be evaluated at different levels where different facts are available for grounding and therefore have different truth values.

Let me provide a more precise comparison with Kripke's theory in order to make it clear exactly what is responsible for the failure of intersubstitutivity. The present theory is the result of changing four things: Quotes are added, the falsity predicate is added, the undefinedness predicate is added, and sentences are "given up" at levels below the fixed point. There is a straightforward way to extend Kripke's semantics to a language where quotes and the falsity predicate are added. In the resulting theory, intersubstitutivity *salva veritate* of co-referential terms would hold. Adding the undefinedness predicate but stopping short of the last change, i.e. interpreting this predicate with an empty extension and an empty anti-extension (to use Kripke's terms), would also not undermine the principle. (Let us call this "Kripke's extended theory".) Only when the last change is made does that happen.

To locate the source of the failure of the principle with more precision, note that the theory could be changed to give intersubstitutivity just by amending the rules for giving up sentences slightly. This could be accomplished simply by stipulating that when a sentence becomes undefined, so

[12]Context theory has been formulated and defended by several authors including Burge [5], Skyrms [18], Goldstein [11] and [12], Simmons [17] and Glanzberg [10].

do all sentences which are identical with it except for different terms with the same reference. As this is simply a matter of giving up more sentences, it does not affect the consistency and fixed point results. Some expressive strength would be lost. For instance, it would no longer be possible to express the fact that "The Strengthened Liar is not true" is not true. However, the idea of assigning the truth value of undefined along the way, as it were, still adds considerable expressibility compared with Kripke's theory. It would still be possible to express the fact that The Strengthened Liar is undefined.

As Kripke's extended theory has the same syntax as the present theory, it facilitates another comparison that is difficult to make directly with Kripke's original theory, namely one about "how much" is made true and false. It is very simple: Every sentence that is true (false) in Kripke's extended theory is true (false) in the present theory. Here is a proof sketch: If sentence s is made true or false in Kripke's extended theory, then at each level below the one where it is given that truth value, at least one of the sentences on which it depends is given a truth value. So E2 ensures that s is not given up. Therefore, the conclusion can be reached by induction, using monotonicity.

9. Tarskian schemata

We get different versions of the Tarskian T-sentence depending on whether the universal quantifier and the bi-conditional are "internal" or "external", i.e. in the object-language or in the meta-language. The external version looks like this: *For all sentences s and terms t such that $I(t) = s$, $T(t)$ is true iff s is true.* The intermediate version, in which the bi-conditional is internal and the quantifier external, can be formulated as follows: *For all sentences s and terms t such that $I(t) = s$, $T(t) \leftrightarrow s$ is true.* Let P be a unary predicate such that $I(P)$ is the set of all sentences of the form $T(t) \leftrightarrow s$ where s is a sentence and t is a term such that $I(t) = s$. Then the internal T-schema, in which both the bi-conditional and the quantifier are in the object-language, giving an object-language generalization about the whole semantics, can be formulated as $\forall v(P(v) \rightarrow T(v))$ *is true.*[13]

Measuring expressive strength with this yardstick, Kripke's theory only earns a score of 1 out of 3; the external version holds but the intermediate and the internal both fail. The Liar is a counter-example to both. But it should not be: The Liar does not make one side true and the other not true, so it should be possible to ground the truth of the bi-conditional in the undefinedness of the Liar. The result for the present theory is different and, I will argue, better, but not perfect.

[13] Although $\forall v(P(v) \rightarrow T(v))$ is a closed sentence, it will incorrectly be referred to as a schema in the interest of simple terminology.

We have to distinguish between more versions of the Tarskian schema. In addition to the difference between external, intermediate and internal versions, there is, with the addition of more truth value predicates, a version with the truth predicate, a version with the falsity predicate and a version with the undefinedness predicate. On top of this, there is a distinction to be made along a third dimension, namely between schemata that are about all names for sentences and schemata that are only about quotation-names.

In this theory, the external and the intermediate Tarskian quotation-names T-, F- and U-schemata hold. That is the content of the next two theorems. The internal T-, F- and U-schemata will be discussed below.

THEOREM 9.1 (External Tarskian quotation-names T-, F- and U-schemata). *For every model and every sentence s the following holds:*

- $[\![T(\ulcorner s \urcorner)]\!] = \top$ *iff* $[\![s]\!] = \top$
- $[\![F(\ulcorner s \urcorner)]\!] = \top$ *iff* $[\![s]\!] = \bot$
- $[\![U(\ulcorner s \urcorner)]\!] = \top$ *iff* $[\![s]\!] = +$

PROOF. Let a sentence s be given. The three bullets can be proved analogously, so we just take the first: Assume $[\![s]\!] = \top$. Let α be the smallest ordinal such that $[\![s]\!]^\alpha = \top$. For all ordinals $\beta < \alpha$, we have $[\![s]\!]^\beta = \mathsf{I}$ and hence by TE9 and E4 for all ordinals $\beta \leq \alpha$, $[\![T(\ulcorner s \urcorner)]\!]^\beta = \mathsf{I}$. It follows that $[\![T(\ulcorner s \urcorner)]\!] = [\![T(\ulcorner s \urcorner)]\!]^{\alpha+1} = \top$. The opposite direction follows directly from TE9. □

Note that this does not mean that s and $T(\ulcorner s \urcorner)$ are interchangeable – they do not necessarily have the same truth value. When $T(\ulcorner s \urcorner)$ is false, s can be undefined. The predicates T, F, and U applied to quotes function as determiners; while s can be true, false or undefined, $T(\ulcorner s \urcorner)$, $F(\ulcorner s \urcorner)$ and $U(\ulcorner s \urcorner)$ always have a proper truth value.

THEOREM 9.2 (Intermediate Tarskian quotation-names T-, F-, and U-schemata). *For every model and every sentence s the following holds:*

- $[\![T(\ulcorner s \urcorner) \leftrightarrow s]\!] = \top$
- $[\![F(\ulcorner s \urcorner) \leftrightarrow (\neg s \wedge \neg U(\ulcorner s \urcorner))]\!] = \top$
- $[\![U(\ulcorner s \urcorner) \leftrightarrow (\neg T(\ulcorner s \urcorner) \wedge \neg F(\ulcorner s \urcorner))]\!] = \top$

PROOF. I give the proof for the second bullet: With the exceptions of s and $\neg s$, all the sub-formulas of $F(\ulcorner s \urcorner) \leftrightarrow (\neg s \wedge \neg U(\ulcorner s \urcorner))$ quote s and are hence evaluated at higher levels than s. $\neg s$ will either get its truth value as a function of the truth value of s or become undefined at the same level as s. $\neg s$ cannot become undefined prior to s getting a truth value because for every evaluation $\mathcal{E}$ containing neither s nor $\neg s$, we have $\overline{R}_\mathcal{E}(s) \subseteq \overline{R}_\mathcal{E}(\neg s)$, so

each of the conditions E2, E3 and E4 holds for $\neg s$ if it holds for s. This leaves four possibilities for the combination of truth values for s and $\neg s$ (mentioned in that order): 1) $\top$ and $\bot$, 2) $\bot$ and $\top$, 3) $+$ and $\top$ and 4) $+$ and $+$. It is easy to check that each possibility results in $F(\ulcorner s \urcorner) \leftrightarrow (\neg s \wedge \neg U(\ulcorner s \urcorner))$ getting the value $\top$. □

In this theory, none of the all-names schemata hold. A counter-example to the external T-schema is given by the sentences $T(c)$ and $U(\ulcorner T(c) \urcorner)$ where c is a constant denoting the latter sentence. The first sentence becomes undefined and the second true. Instead of the external Tarskian all-names T-, F- and U-schemata, we have this weaker theorem:

THEOREM 9.3. *For every model $\mathfrak{M} = (D, I)$, sentence s, and constant c such that $I(c) = s$ the following holds:*

- *If $[\![T(c)]\!] = \top$, then $[\![s]\!] = \top$. If $[\![s]\!] = \top$, then $[\![T(c)]\!] \in \{\top, +\}$.*
- *If $[\![F(c)]\!] = \top$, then $[\![s]\!] = \bot$. If $[\![s]\!] = \bot$, then $[\![F(c)]\!] \in \{\top, +\}$.*
- *If $[\![U(c)]\!] = \top$, then $[\![s]\!] = +$. If $[\![s]\!] = +$, then $[\![U(c)]\!] \in \{\top, +\}$.*

PROOF. First bullet: The first implication follows directly from TE9. To prove the second implication, assume $[\![s]\!] = \top$. $T(c)$ either receives a truth value at a higher level than s or at the same or a lower level. In the first case, we have $[\![T(c)]\!] = \top$ by TE9. In the second case, none of the bullets of TE9 are satisfied at the level where $T(c)$ gets a value, so it follows that $[\![T(c)]\!] = +$.

For the two other bullets, just replace "$T(c)$" with "$F(c)$" and "$U(c)$" respectively and "TE9" with "TE10" and "TE11" respectively. □

The internal quotation-names T-schema can be formulated as $\forall v(P(v) \rightarrow T(v))$, where P is a unary predicate such that $I(P)$ is the set of all sentences of the form $s \leftrightarrow T(\ulcorner s \urcorner)$ where s is a sentence. In the present theory, this sentence becomes undefined. The problem is that $\forall v(P(v) \rightarrow T(v))$ can only become true after all instances of $s \leftrightarrow T(\ulcorner s \urcorner)$ have been made true. And one of these instances is the one where s is $\forall v(P(v) \rightarrow T(v))$. The same holds *mutatis mutandis* for the internal F- and U-schemata. More generally, this theory shares the problem with Kripke's that intuitively true generalizations about the whole semantics are not made true by the theory.

That concludes the factual account of what versions of the Tarskian schema that hold and fail to hold in the theory, and we can turn to the discussion of whether the result is reasonable. For simplicity, let us restrict the discussion to the schemata for truth and ignore falsity and undefinedness. We have six different Tarskian schemata for truth. There is the distinction between external, intermediate and internal, and for each of these there is a

schema which is about all names for sentences versus a schema which is only about quotation-names. Tarski's condition of adequacy can be stated like this: Any instance of any of the schemata that can be formulated in a given language should be true according to an adequate theory of that language.

Is the adequacy condition reasonable? An argument for a positive answer is given by Field. He notes that among the primary purposes of the truth predicate is that it is a means of expressing agreement and disagreement and "a device of quantification". For the former, he gives this example:

> *Jones makes some complicated bunch of claims that I agree with, and instead of expressing agreement by repeating the whole thing I say "What he said is true."* [8, p. 138]

If the truth predicate is to serve its purpose here, Jones' "complicated bunch of claims" must be true if and only if Field's utterance "What he said is true" is true. And the latter is precisely the truth predicate applied to a name (or rather description) of Jones' assertions. Ergo, the external T-schema should hold. The same conclusion follows from Field's example of the use of the truth predicate as a device of quantification more generally than its use of expressing agreement and disagreement:

> *Suppose I can't remember exactly what was in the Conyers report on the 2004 election, but say*
> *(1) If everything that the Conyers report says is true then the 2004 election was stolen.*
> *Suppose that what the Conyers report says is* $A_1, \ldots, A_n$. *Then relative to this last supposition, (1) better be equivalent to*
> *(2) If* A_1 *and . . . and* A_n *then the 2004 election was stolen.* [8, p. 210]

Actually, Field concludes that $T(t)$ and s should have the same truth value (when $I(t) = s$), but the argument does not support this stronger conclusion. If s is undefined then it is quite reasonable that $T(t)$ is false. The argument can at most establish two things, the conjunction of which is weaker than Field's conclusion: First, that s should be true if and only if $T(t)$ is, and second, that the truth values of s and $T(t)$, when being the truth value of the antecedent of a conditional, should have the same effect on the truth value of the conditional. Or more generally: The truth values of s and $T(t)$ when being the truth value of some proper component of a composite sentence (and not in the scope of further truth value predicates) should have the same effect on the truth value of that sentence.

This implies that the partial success of the theory as expressed in theorems 9.1 and 9.2 is genuine; we should not have interchangeability of $T(\ulcorner s \urcorner)$ and s instead of the first bullet in theorem 9.1, and the object language bi-conditional (see section 4) is the right one to use in the formulation of

theorem 9.2 and should not be replaced with equivalence. It also speaks to the reasonableness of letting falsehood and undefinedness play the same role when being "input" to connectives and quantifiers, as it indeed was (same section).

Let us turn to the all-names schemata. As explained they are not validated by the theory. That is an apparent shortcoming which I will argue is not genuinely so. With the metaphor of the writer of the two books at hand, it is not too hard to accept that the three premises of the semantical paradoxes, classical logic, unrestricted validity of the T-schemata and semantic closure, cannot hold jointly, as it is when one is first confronted with them. The assumption of classical logic corresponds to assuming that all facts about the contents of the books can be taken as given before they are written. The assumption of the unrestricted validity of the T-schemata corresponds to a belief that reality (the right-hand side of the T-schemata) is completely represented in the books (the left-hand side of the T-schemata). And the assumption of semantic closure can be interpreted as the belief that this representation is included in the reality. Taken together these three assumptions amount to the belief that a complete picture of all of reality exists as a proper part of that same reality; that it is possible to have a view from nowhere on the totality of reality and at the same time expect the picture of this view to exist inside the reality. When the premises are seen in this light, the need to give up on the conjunction of the three premises seems much less of a loss.

The same point can be formulated in a different way. The work of the writer can be understood as consisting of observing and recording. He observes the fact that grass is green and records this observation by declaring "Grass is green" true. For it to be possible that all of the three premises were true, it would have to be possible to have a world that was stable with respect to any such observe-and-record act, even though everything in the world, including all records of observations, were observable. The failure of the conjunction of the three premises can be interpreted as the existence in semantics, as in physics, of an "observer effect".

The failure of the all-names schemata is intuitively acceptable despite Field's argument, that Jones' claims should be true if and only if Field's sentence is. I submit that this conclusion is only reasonable if it is the case that, so to speak, the truth values of Jones' sentences are a given from the perspective of Field's sentence, i.e. if they do not depend on that very sentence. If Field's act of "observing and recording" affects the observed, he must accept the risk that his records may not reflect it accurately.

And yet we have, in virtue of the validation of the external quotation name schema, what could be called the better part of semantic closure: Ev-

ery individual semantic fact of the language can be expressed in the language (setting aside the issue of semantic facts about other things than truth values). By "individual", I mean semantic facts which are each concerned with just a single sentence. As the external schemata do not hold for all names, but do hold for quotation-names, it is not possible to express every individual semantic fact with *any* expression that one may naïvely assume could be used, such as the sentence "This sentence is not true" to express that that very sentence is not true, but for every individual semantic fact there will be *some* sentence in the language to express it, namely one using a quotation: *"This sentence is not true" is not true.*

But nevertheless this brings us to the genuine shortcoming of the theory. For having semantic closure merely with respect to individual semantic facts cannot be defended with reference to the nature of the work of the writer. The dependency relation imposed on semantics does not justify the failure of the internal quotation-names schema. For the truth of the internal quotation-names schema is something that the writer would be able to realise prior to assigning truth values to all sentences. He could do that by reflecting on the structure of the semantics, for example by going through the reasoning of the proof of theorem 9.2, instead of having to rely on inspection of each and every sentence. So in this respect, the theory is inadequate. However, to solve this problem we have to allow grounding in intensional properties instead of just in extensional properties, and that is also a subject I must postpone to another paper.

10. An objection

A possible objection to the theory is that it seems to sacrifice a lot. In addition to classical logic, both the full validity of the external T-schema and the principle of intersubstitutivity of coextensional terms are given up. One may question why one should take this theory seriously when there are other theories that sacrifice less. For instance the Kripke-Feferman theory [7] invalidates the T-schema but maintains both classical logic and intersubstitutivity, while Skyrms [18] accomplices the opposite, i.e. he retains the T-schema and pays with just classical logic and intersubstitutivity.

The idea behind such an objection would seem to be that we can define a partial order $<$ on the theories of truth by having theory A$<$theory B, if A rejects all the intuitive principles that B rejects plus some more, and that a theory A is clearly wrong if there exists a theory B such that A$<$B. For we should seek to minimise the discrepancy between our theory of truth and our pre-theoretic ideas about truth. The ideal is a consistent theory which combines classical logic, the unrestricted T-schema, semantic closure, etc.,

and it is just a damn shame that it doesn't exist. Other theories should be measured by their proximity to that ideal.

I see two ways to respond to this objection. This is the first: The minimization strategy is in a sense based on a refusal to really accept Tarski's theorem. The idea is that a theory of truth *should* use classical logic, *should* validate the unrestricted T-schema, *should* produce a semantically closed language and *should* be consistent. We should get as close as possible to the unobtainable ideal because the unobtainable ideal is the only theory that is not wrong in at least one aspect. But this is absurd. We should not strive for the impossible, but find out why it is impossible – why the apparent ideal is unobtainable. Behind a theory of truth there should be a coherent story, which in addition to motivating the theory explains why the principles that it rejects are not correct. Such a story might well imply that several naïve principles have to be rejected even if it is technically possible to reject just some of those.

This is my story: There are words and there is the world, and when the former correspond to the latter, there is truth. Yet, the words are in the world, so in some cases the part of the world that some (combination of) words attempts to correspond to cannot be determinate independently of those very words themselves. That is why classical logic, in particular *tertium non datur*, fails. But that failure is itself a fact in the world to which other words can correspond and when they can, they have a proper truth value. In some cases one sentence cannot correspond to a fact (because the fact depends on the truth value of that same sentence) while another sentence can (because the fact does not depend on *that* sentence), even though the two sentences are making the same claim. That is why the all-names T-schema and substitutivity fail.

The second way of responding is to point out that the same objection could be raised against Kripke's theory and is unreasonable in that case. Tarski rejects semantic closure. Kripke rejects classical logic, the intermediate and internal T-schemata *and* semantic closure. So it should be obvious that Kripke's theory is inferior to Tarski's. It is not, however, for Kripke has actually not sacrificed anything compared to Tarski. For every sentence that can be formulated in Tarski's language has the same truth value in Kripke's language, so in particular classical logic holds for the T-free fragment of the latter. The choice between Tarski's and Kripke's theories is not a choice between classical logic and a non-standard logic for a common class of sentences. It is a choice between not formulating these sentences at all on the one hand, and formulating them and taking the risk that they may be neither true nor false on the other. And it is only the sentences that add the extra expressive strength that are counterexamples to those principles that

do not hold in Kripke's theory but do in Tarski's. The same is the case when the theory set forth in this paper is compared to Kripke's; the loss of the all-names external T-schema and of intersubstitutivity is only due to sentences which are undefined in Kripke's theory but get a proper truth value in this theory and to sentences which can't even be formulated in Kripke's, cf. section 8. In an important sense nothing has been sacrificed compared to Kripke's theory. Doing the score simply by counting rejected principles is misleading.

Acknowledgements. For their valuable comments to drafts of this paper I wish to thank Vincent Hendricks, Crispin Wright, Aaron Cotnoir, Elia Zardini, Andreas Fjellstad, Øystein Linnebo, the anonymous referees and the participants of at the colloquium "PhD's in Logic II" at the University of Tilburg on the 19th of February 2010, in particular Leon Horstein and Stefan Wintein.

References

[1] Austin, J. L., 'Truth', *Proceedings of the Aristotelian Society, Supplementary Volumes*, 24 (1950), 111–128.

[2] Beall, J. C., (ed.) *Revenge of the Liar: new essays on the paradox*, Oxford University Press, 2007.

[3] Belnap, N. D., 'A useful four-valued logic', in J. Michael Dunn, and George Epstein, (eds.), *Modern Uses of Multiple-Valued Logic*, D. Reidel Publishing Company, 1977.

[4] Bolander, Thomas, *Logical theories for agent introspection*, Ph.D. thesis, Technical University of Denmark, 2003.

[5] Burge, Tyler, 'Semantical paradox', *The Journal of Philosophy*, 76 (1979), 169–198.

[6] Dunn, J. Michael, 'Intuitive semantics for first-degree entailments and 'coupled trees'', *Philosophical Studies*, 29 (1976), 149–168.

[7] Feferman, Solomon, 'Towards useful type-free theories I', *The Journal of Symbolic Logic*, 49 (1984), 75–111.

[8] Field, Hartry, *Saving truth from paradox*, Oxford University Press, 2008.

[9] Fitting, Melvin, 'Notes on the mathematical aspects of Kripke's theory of truth', *Notre Dame Journal of Formal Logic*, 27 (1986), 75–88.

[10] Glanzberg, Michael, 'The liar in context', *Philosophical Studies*, 103 (2001), 217–251.

[11] Goldstein, Laurence, "This statement is not true' is not true', *Analysis*, 52 (1992), 1–5.

[12] Goldstein, Laurence, 'Truth-bearers and the Liar – a reply to Alan Weir', *Analysis*, 61 (2001), 115–26.

[13] Kripke, Saul, 'Outline of a theory of truth', *The Journal of Philosophy*, 72 (1975), 690–716.

[14] Leitgeb, Hannes, 'What truth depends on', *Journal of Philosophical Logic*, 34 (2005), 155–192.

[15] Priest, Graham, *Beyond the Limits of Thought*, second edn., Oxford University Press, 2002.

[16] Schaffer, Jonathan, 'On what grounds what', in David Chalmers, David Manley, and Ryan Wasserman, (eds.), *Metametaphysics: New Essays on the Foundations of Ontology*, Oxford University Press, 2009.

[17] Simmons, Keith, *Universality and the Liar*, Cambridge University Press, 1993.

[18] Skyrms, Brian, 'Intensional aspects of semantical self-reference', in Robert L. Martin, (ed.), *Recent essays on truth and the liar paradox*, Oxford University Press, 1984.

[19] Tarski, Alfred, 'The semantic conception of truth', *International Phenomenological Society*, 4 (1944), 341–376.

[20] Yablo, Stephen, 'Grounding, dependence, and paradox', *Journal of Philosophical Logic*, 11 (1982), 117–137.

[21] Yablo, Stephen, 'Paradox without self-reference', *Analysis*, 53 (1993), 251–252.