Restricted truth predicates in first-order logic

Thomas Bolander

1 Introduction

It is well-known that there exist consistent first-order theories that become inconsistent when we add Tarski's schema T. This is Tarski's Theorem. To avoid the inconsistency result, one can restrict Tarski's schema in different ways. In our paper we restrict Tarski's schema T by only instantiating the schema with a proper subset of the set of all sentences. We prove several results concerning the sets of sentences M for which Tarski's schema T instantiated with the sentences of M is relatively consistent with any first-order theory.

Let $\mathcal{L}$ be any first-order language containing the one-place predicate symbol T (intended to denote truth). Let M be a subset of the set of sentences of $\mathcal{L}$. By the **truth predicate** over M we understand the instances over M of Tarski's schema T, that is, the theory

$$\{\varphi \leftrightarrow T(\ulcorner\varphi\urcorner) \mid \varphi \in M\},$$

where $\ulcorner\cdot\urcorner$ is some suitable coding scheme. Such theories are also called **restricted truth predicates**. Let $\mathcal{L}^-$ denote the first-order language $\mathcal{L}$ with the predicate symbol T removed. The truth predicate over M is called **universally consistent** if it is relatively consistent with any theory in $\mathcal{L}^-$.[1] We are interested in knowing for which sets of sentences M the truth predicate over M is universally consistent. We already know that this is not the case when $M = \mathcal{L}$ (Tarski's Theorem. See Tarski 1956). Based on this negative result, our main question become: *What are the maximal M for which the truth predicate over M is universally consistent?*[2] Our tools to investigate these problems will be *semantic functionals* and *dependence relations*. These are introduced in Section 4 and 5. In Section 6 and 7 we prove several results concerning universally consistent truth predicates. We will be generalizing previous results by Perlis 1985, Feferman 1984, and des Revières and Levesque 1986. Our main result is Theorem 8 (p. 7). Proofs can be found in the appendix.

2 Motivation

The problem of determining for which sets M the truth predicate over M is universally consistent has an interest in at least the following three fields of research:

(i) Philosophy—especially regarding *truth definitions* in the philosophy of language.

(ii) Computer science—especially regarding *introspective agents* in artificial intelligence.

[1] That is, if U is any consistent first-order theory in $\mathcal{L}^-$ then U extended with the truth predicate over M is also consistent.

[2] Vann McGee considers a closely related problem in McGee 1992. But McGee finds maximal truth predicates that are relatively consistent with particular theories, and not maximal truth predicates that are relatively consistent with *any* theory.

(iii) Mathematics—especially regarding *axiomatic set theories* in mathematical logic.

In *philosophy*, the naive theory of truth requires that the truth predicate, formalized by T, satisfies each of the equivalences

$$\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \tag{1}$$

where φ range over the sentences of the language considered. But the naive theory of truth is challenged by the fact that the theory

$$\{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \text{ is a sentence}\}$$

is *not* universally consistent. That is, not every consistent theory formalized in first-order logic can consistently be extended with a truth predicate T satisfying (1) above.

In *computer science*, one of the important problems is the question of how to implement introspection, or self-reflection, in artificial intelligence agents. It turns out that the problem of constructing an agent which possesses complete and correct introspective abilities is—in a first-order logical setting—equivalent to the problem of having a first-order theory in which all instances of (1) are theorems. Thus, in a first-order logical setting, the problem of complete introspection of agents is more or less identical to the problem of truth definitions. This is actually not too surprising: both the construction of truth predicates and obtaining complete introspection amounts to constructing a theory a part of which contains a complete representation of the theory itself (in our case the part that concerns the T predicate).

In *mathematics*, restricted truth predicates are relevant to axiomatic set theory amongst others. By the **abstraction principle** over a set M of sentences of a suitable first-order language $\mathcal{L}$ we understand the theory consisting of all sentences

$$\forall x\,(x \in \{y \mid \varphi(y)\} \leftrightarrow \varphi(x))$$

where $\varphi(x) \in M$. The naive theory of sets is the abstraction principle over all sentences of $\mathcal{L}$. The naive theory of sets is inconsistent (this is shown by a formalization of Russell's paradox). Thus, in axiomatic set theory, one needs to restrict the abstraction principle in one way or another to regain the essential consistency. This can for instance be done by only considering the abstraction principle over some subset M of the set of sentences of $\mathcal{L}$. The truth predicate over M,

$$\{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \in M\}\,,$$

is interpretable in the abstraction principle over M.[3] Under certain conditions, the abstraction principle over M is also interpretable in the truth predicate over M. Thus, results about the consistency of theories containing restricted truth predicates also gives information on the consistency of restricted abstraction principles.

The above discussion shows that proving consistency results for restricted truth predicates can be of importance to all of the three fields of research mentioned in the beginning of this section. Besides and beneath these motivations for studying restricted truth predicates lies an attempt to understand and model the essential features of self-reflection and self-reference.

[3]This is not hard to show. See e.g. Bealer 1982.

3 Basic Definitions and Conventions

We use $\mathcal{L}, \mathcal{L}'$, etc. to range over languages of first-order predicate logic containing the predicate symbol T. For every first-order language $\mathcal{L}$, we let $\mathcal{L}^-$ denote the language obtained by removing the predicate symbol T from $\mathcal{L}$. We use U, V, U', V', etc. to range over theories of first-order predicate logic. Predicate logic is here taken to mean predicate logic *with identity*. This means that: 1) all languages contain the identity symbol $=$; 2) all theories contain the axioms of equality as part of their logical axioms; 3) all models are normal.[4] Every first-order language $\mathcal{L}$ is identified with its set of sentences. We assume the connectives of first-order logic to be $\neg, \wedge$ and $\vee$. When using $\rightarrow$ and $\leftrightarrow$ in a formula this should be read as an abbreviation for the corresponding formula using only $\neg, \wedge$ and $\vee$.

To make things technically simpler, we will make some further assumptions on the languages and theories to be used. We will assume that every language considered contains 0 as its only constant symbol and the one-place function symbol s at its only function symbol. For theories with equality this does not restrict the generality of the results, since all constant and function symbols in theories with equality can be dispensed with by introducing new predicate symbols instead (Mendelson 1997). We will furthermore assume that every theory U considered satisfy

$$U \vdash \bar{n} \neq \bar{m} \quad \text{when } n, m \in \mathbb{N} \text{ with } n \neq m.^5$$

Throughout the paper we will be using a fixed Gödel coding $\ulcorner \cdot \urcorner$ that maps any formula of any first-order language into an element of $\mathbb{N}$. The only property we will need this coding to satisfy is that $\ulcorner \varphi(\ulcorner \psi \urcorner) \urcorner > \ulcorner \psi \urcorner$ for all φ, ψ. For every formula φ we identify $\ulcorner \varphi \urcorner$ with the numeral $\bar{n}$ denoting this Gödel number. Thus in every theory, for every formula φ, there is a term $\ulcorner \varphi \urcorner$ that we can use to denote φ.

The set of closed terms (ground terms) of $\mathcal{L}$ is denoted $\mathrm{Terms}(\mathcal{L})$. The set of atomary sentences (closed atomary formulas) is denoted $\mathrm{Atoms}(\mathcal{L})$ and the set of well-formed formulas is denoted $\mathrm{wff}(\mathcal{L})$.

4 Semantic Functionals

To prove our universal consistency results we will introduce a general method that can be used to prove many such consistency results.

Let a first-order language $\mathcal{L}$ be given. We use t to denote the truth-value "true" and f to denote the truth-value "false". $\{t, f\}_\perp$ denotes the set $\{t, f, \perp\}$. We use mappings $\mathcal{L} \rightarrow \{t, f\}_\perp$ to represent *partial* functions from $\mathcal{L}$ into $\{t, f\}$, where the value $\perp$ means "undefined". We expect the reader to be familiar with the basics of partial functions treated in this way (otherwise consult e.g. Stoltenberg-Hansen et al. 1994). By a **semantic functional** we then understand any map

$$F : (\mathcal{L} \rightarrow \{t, f\}_\perp) \rightarrow (\mathcal{L} \rightarrow \{t, f\}_\perp),$$

[4] A model is called **normal** if the predicate symbol $=$ is interpreted as the identity on the domain of the model.

[5] As usual, we take $\bar{n}$ for all $n \in \mathbb{N}$ to denote the term $s(s(\ldots s(0)\ldots))$ with n occurrences of s. $\bar{n}$ is the **numeral** representing n.

that is, any map that takes a partial truth-value assignment $[\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp$ and returns a new partial truth-value assignment. Such a partial truth-value assignment is called a **fixed point** of a F if $F([\![\cdot]\!]) = [\![\cdot]\!]$, that is, if $F([\![\cdot]\!])(\varphi) = [\![\varphi]\!]$ for all $\varphi \in \mathcal{L}$. It is called a **total fixed point** if it is a total function, that is, if $[\![\varphi]\!] \in \{t, f\}$ for all $\varphi \in \mathcal{L}$. When we talk about semantic functionals in the following we will often just call them functionals.

We define a class of semantic functionals that are of particular interest in our case. For any first-order language $\mathcal{L}$ and any set of sentences $M \subseteq \mathcal{L}$ we define the functional $F_{\mathcal{L},M}$ by

$$F_{\mathcal{L},M} : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$$

by, for all sentences $\varphi, \psi, \forall x \alpha(x) \in \mathcal{L}$,

$$F_{\mathcal{L},M}([\![\cdot]\!])(\varphi \wedge \psi) = \begin{cases} t & \text{if } [\![\varphi]\!] = [\![\psi]\!] = t \\ \perp & \text{if } [\![\varphi]\!] = \perp \text{ or } [\![\psi]\!] = \perp \\ f & \text{otherwise} \end{cases} \tag{2}$$

$$F_{\mathcal{L},M}([\![\cdot]\!])(\neg\varphi) = \begin{cases} f & \text{if } [\![\varphi]\!] = \text{t} \\ t & \text{if } [\![\varphi]\!] = \text{f} \\ \perp & \text{otherwise} \end{cases} \tag{3}$$

$$F_{\mathcal{L},M}([\![\cdot]\!])(\forall x \alpha(x)) = \begin{cases} t & \text{if } [\![\alpha(\tau)]\!] = t \text{ for all terms } \tau \text{ in } \mathcal{L} \\ \perp & \text{if } [\![\alpha(\tau)]\!] = \perp \text{ for some term } \tau \text{ in } \mathcal{L} \\ f & \text{otherwise} \end{cases} \tag{4}$$

$$F_{\mathcal{L},M}([\![\cdot]\!])(T(\ulcorner\varphi\urcorner)) = [\![\varphi]\!] \text{ for all } \varphi \in M \tag{5}$$

The first three clauses above are recognized to correspond to the truth conditions of Kleene's weak three-valued logic with $\perp$ as the third value. The fourth clause ensure that in any fixed point $[\![\cdot]\!]$ of $F_{\mathcal{L},M}$ we will have

$$[\![T(\ulcorner\varphi\urcorner)]\!] = [\![\varphi]\!] \text{ for all } \varphi \in M, \tag{6}$$

that is, for all $\varphi \in \mathcal{M}$, $[\![\cdot]\!]$ assigns the same truth value to $T(\ulcorner\varphi\urcorner)$ as to φ. This is the property that will allow us to prove universal consistency results using these functionals.

Let $F : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$ be a functional. The set of **initial elements** of F, denoted $init(F)$, is defined as

$$\text{init}(F) = \{\varphi \in \mathcal{L} \mid F([\![\cdot]\!])(\varphi) = \perp \text{ for all } [\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp\}.$$

For $F_{\mathcal{L},M}$ defined above we have

$$\text{init}(F_{\mathcal{L},M}) = \{\varphi \in \mathcal{L} \mid \varphi \text{ is an atomary sentence not on the form } T(\ulcorner\psi\urcorner) \text{ for any } \psi \in M\}.$$

That is, $\text{init}(F_{\mathcal{L},M})$ is the set of sentences that are not covered by any of the defining clauses for $F_{\mathcal{L},M}$. By an **initial extension** of F we understand any functional G that extends F as a partial function and satisfies:

$$G([\![\cdot]\!])(\varphi) \in \{t, f\} \text{ for all } \varphi \in \text{init}(F) \text{ and all } [\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp.$$

The connection between our semantic functionals and universal consistency of truth predicates is revealed by the following important lemma.

Lemma 1. *Let $M \subseteq \mathcal{L}$. If for every first-order language $\mathcal{L}' \supseteq \mathcal{L}$, every initial extension of $F_{\mathcal{L}',M}$ has a total fixed point then the theory*

$$\{\varphi \leftrightarrow T(\ulcorner\varphi\urcorner) \mid \varphi \in M\}$$

is universally consistent.

Thus the problem now becomes to prove the existence of total fixed points of semantic functionals. For this we need the notion of a dependence relation.

5 Dependence Relations

Consider again (2), the first clause in the definition of $F_{\mathcal{L},M}$. This is the clause for the conjunction $\varphi \wedge \psi$. Since both $[\![\varphi]\!]$ and $[\![\psi]\!]$ occur on the right hand side of the clause we can infer that

> "$\varphi \wedge \psi$ is semantically dependent on φ and ψ"

or that

> "to determine the semantic value of $\varphi \wedge \psi$ we first need to determine the semantic values of both φ and ψ".

Similarly, from the clause

$$F_{\mathcal{L},M}([\![\cdot]\!])(T(\ulcorner\varphi\urcorner)) = [\![\varphi]\!] \text{ for all } \varphi \in M$$

we can infer that

> "$T(\ulcorner\varphi\urcorner)$ is semantically dependent on φ (when $\varphi \in M$)".

The *semantical dependence* that is hereby expressed in the semantic functionals can be represented by *dependence relations*. A dependence relation is a binary relation on the set of sentences in which φ is related to ψ iff φ is semantically dependent on ψ, that is, iff ψ appears on the right-hand side of the clause for φ. More formally,

Definition 2. *Let $F : (\mathcal{L} \to \{t, f\}_\perp) \to (\mathcal{L} \to \{t, f\}_\perp)$ be a semantic functional. For each $\varphi \in \mathcal{L}$, we define the **dependence set** for φ (wrt. F) as the least set M such that*

$$\text{for all } [\![\cdot]\!] : \mathcal{L} \to \{t, f\}_\perp, \text{ if } \mathrm{dom}([\![\cdot]\!]) \supseteq M \text{ then } \varphi \in \mathrm{dom}(F([\![\cdot]\!])).^6$$

*If such a least set M does not exist, we let the dependence set for φ be $\emptyset$. The **dependence relation** of F is now defined as the binary relation D on $\mathcal{L}$ given by*

$$\varphi D \psi \quad \Leftrightarrow \quad \psi \text{ is in the dependence set of } \varphi \text{ (wrt. } F\text{).}$$

*When $\varphi D \psi$ we say that φ **depends** on ψ.*

[6] For every function $f : \mathcal{L} \to \{t, f\}_\perp$, $\mathrm{dom}(f)$ is the set of $\varphi \in \mathcal{L}$ for which $f(\varphi) \neq \perp$.

6 Well-founded Truth Predicates

We will now use Lemma 1 to prove our first universal consistency result.

A binary relation D is called **conversely well-founded** is there is no infinite sequence of elements $e_1, e_2, e_3, \ldots$ such that

$$e_1 D e_2 D e_3 \cdots .$$

When we are given a semantic functional for which the dependence relation is conversely well-founded we find ourselves in a very advantageous situation:

Lemma 3. *Let $M \subseteq \mathcal{L}$. If the dependence relation of $F_{\mathcal{L},M}$ is conversely well-founded then any initial extension of $F_{\mathcal{L},M}$ has a total fixed point.*

Combining this result with Lemma 1 we immediately get

Lemma 4. *Let $M \subseteq \mathcal{L}$. If the dependence relation of $F_{\mathcal{L}',M}$ is conversely well-founded for every $\mathcal{L}' \supseteq \mathcal{L}$ then the theory*

$$\{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \in M\}$$

is universally consistent.

To know that the truth predicate over a set M is universally consistent we thus only need to know that the dependence relation of $F_{\mathcal{L}',M}$ is conversely well-founded. For every language $\mathcal{L}$, there is actually a greatest set for which $F_{\mathcal{L},M}$ is conversely well-founded, as the following lemma shows.

Lemma 5. *Let $\mathcal{L}$ be a first-order language. Let $M \subseteq \mathcal{L}$ be given by*

$$M = \{\varphi \in \mathcal{L} \mid \varphi \textit{ does not contain } T(x) \textit{ as a subformula}\} .$$

M is the greatest subset of $\mathcal{L}$ for which the dependence relation of $F_{\mathcal{L},M}$ is conversely well-founded.

This lemma together with Lemma 4 immediately gives us our first universal consistency result.

Theorem 6. *Let $\mathcal{L}$ be a first-order language. The theory*

$$\{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \textit{ does not contain } T(x) \textit{ as a subformula}\}$$

is universally consistent.

This result tells us that we can always safely instantiate Tarski's schema T with sentences such as

$$T(\ulcorner \neg T(\ulcorner on(cat, mat) \urcorner) \urcorner)$$

in which we have nested and/or negated occurrences of T, but not necessarily with sentences such as

$$\forall x \, (P(x) \vee T(x))$$

where we have quantified truth.

Corollary 7. *As special cases of the above theorem we get that both of the following theories are universally consistent:*

(i) $\{\varphi \leftrightarrow T(\ulcorner\varphi\urcorner) \mid \varphi$ *does not contain the predicate symbol* $T\}$.

(ii) $\{\varphi \leftrightarrow T(\ulcorner\varphi\urcorner) \mid \varphi$ *does not contain variables*$\}$.

(i) above is the Tarski idea: we only apply the truth predicate to sentences that do not themselves contain the truth predicate. (ii) gives us a truth predicate that is restricted in only applying to singular statements (that is, "the sentence 'the cat is on the mat' is true" is included but "every sentence is true" is not). Such a truth predicate might be sufficient for many applications to introspective agents.

The result above can actually be immediately strengthened by using a more general type of coding. By a **parametrized coding** in $\mathcal{L}$ we understand an injective map $\ulcorner\cdot\urcorner$ from the formulas of $\mathcal{L}$ into the terms of $\mathcal{L}$ satisfying

(i) For any formula φ in $\mathcal{L}$, $\ulcorner\varphi\urcorner$ has the same free variables as φ.

(ii) For any formula $\varphi(x)$ and any term τ which is free for x in $\varphi(x)$, $\ulcorner\varphi(\tau)\urcorner$ is the term obtained by substituting τ for all free occurrences of x in $\ulcorner\varphi(x)\urcorner$.

Unfortunately, the space does not allow us to prove the existence of such parametrized codings, but the interested reader is referred to Feferman 1984 in which a closely related kind of parametrized coding is constructed. Using a parametrized coding scheme rather than a standard Gödel coding does not change the validity of Theorem 6. This is easily seen from the proof of Theorem 6 given in the appendix. Using a parametrized coding allows us to quantify into arguments of the T predicate as in for instance

$$\exists x T\left(\ulcorner on(cat, x)\urcorner\right).$$

Since such sentences have no occurrence of $T(x)$, the parametrized version of Theorem 6 shows that we can also safely instantiate Tarski's schema T with these. The parametrized version of Theorem 6 is a generalization of the main theorem of des Rivières and Levesque 1986.

7 Positive Truth Predicates

We will now prove universal consistency of an even bigger set of instances of Tarski's schema T.

Theorem 8. *Let* $\mathcal{L}$ *be a first-order language. The theory*

$$\{\varphi \leftrightarrow T(\ulcorner\varphi\urcorner) \mid \varphi \text{ does not contain } T(x) \text{ as a subformula in the scope of } \neg\}$$

is universally consistent.

Note that we are only using the connectives $\neg, \wedge$ and $\vee$, so e.g. the formula

$$\forall x\left(T(x) \rightarrow A(x)\right)$$

will contain $T(x)$ in the scope of $\neg$, since this sentence is an abbreviation for

$$\forall x\,(\neg T(x) \vee A(x))\,.$$

Compared to Theorem 6 this result shows us that it is also safe to instantiate Tarski's schema T with sentences such as

$$\forall x(A(x) \wedge T(x))$$

in which we have occurrences of $T(x)$—as long as these occurrences are not negative. This theorem generalizes results from Perlis 1985 and Feferman 1984, in that their results only prove it safe to instantiate Tarski's schema T with sentences that has no negative occurrence of $T(\tau)$ for *any* term τ. This means that e.g.

$$\neg T(\ulcorner \varphi \urcorner)$$

is excluded from their solution for all φ, while it is included in the one given here.

We cannot get a universally consistent truth predicate much stronger than given by the above theorem. This is seen by the following fact.

Lemma 9. *Let $\mathcal{L}$ be any first-order language. If $M \subseteq \mathcal{L}$ is a set containing a sentence*

$$\forall x\,(A(x) \vee \neg T(x))$$

where $A \neq T$ then

$$\{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \in M\}$$

is not *universally consistent.*

Appendix

We now give proofs for the results stated above. We do not give proofs for Lemma 4, Theorem 6, and Corollary 7 that were all obvious corollaries of preceding results. Furthermore, the space unfortunately does not allow us to give a proof of Theorem 8. But it is a relatively simple corollary of lemmata 1, 3 and 5.

Proof of Lemma 1. Assume that for every first-order language $\mathcal{L}' \supseteq \mathcal{L}$ every initial extension of $F_{\mathcal{L}',M}$ has a total fixed point. Let $A = \{\varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \in M\}$. We have to prove that A is universally consistent, that is, for every consistent theory U in $\mathcal{L}^-$, $U \cup A$ is consistent.[7] Let thus U be any consistent theory in $\mathcal{L}^-$. Then there exists some language $\mathcal{L}'^- \supseteq \mathcal{L}^-$ in which U has a Herbrand model $\mathcal{M}$ (a model in which the domain is the set of closed terms of the language and every closed term is interpreted as itself). Since we have that $U \vdash \bar{n} \neq \bar{m}$ whenever $n \neq m$, this model can be assumed to be normal. We want to expand the model $\mathcal{M}$ into a model $\mathcal{M}'$ of $\mathcal{L}'$ such that $\mathcal{M}'$ becomes a model of $U \cup A$. This would prove $U \cup A$ to be consistent, and thus we would be done.

[7] Note that $U \cup A$ is a theory in $\mathcal{L}$ and thus also contains the axioms of equality for formulas involving the T predicate.

Let G be any initial extension of $F_{\mathcal{L}',M}$ satisfying

$$G([\![\cdot]\!])(\varphi) = \begin{cases} t & \text{if } \varphi \in \text{init}(F_{\mathcal{L}',M}) \cap \mathcal{L}'^{-} \text{ and } \mathcal{M} \models \varphi \\ f & \text{if } \varphi \in \text{init}(F_{\mathcal{L}',M}) \cap \mathcal{L}'^{-} \text{ and } \mathcal{M} \not\models \varphi. \end{cases} \tag{7}$$

By assumption, G has a total fixed point $[\![\cdot]\!]$. The expansion $\mathcal{M}'$ of $\mathcal{M}$ is then defined by letting

$$T^{\mathcal{M}'} = \left\{ \tau \in \text{Terms}(\mathcal{L}') \mid [\![T(\tau)]\!] = t \right\}. \tag{8}$$

Our goal is now to prove the following.

Claim. For all $\varphi \in \mathcal{L}'$, $\mathcal{M}' \models \varphi$ if and only if $[\![\varphi]\!] = t$.

Proof of claim. We will prove this by induction on the syntactic complexity of φ. Assume first that φ is atomary. If φ is an initial element in $\mathcal{L}'^{-}$ then by (7) we have

$$\mathcal{M}' \models \varphi \Leftrightarrow G([\![\cdot]\!])(\varphi) = t \Leftrightarrow [\![\varphi]\!] = t$$

where the last equivalence follows from the fact that $[\![\cdot]\!]$ is a fixed point of G. If φ is atomary but not initial it must be on the form $T(\tau)$ for some $\tau \in \text{Terms}(\mathcal{L}')$. But then by (8) we immediately get

$$\mathcal{M}' \models T(\tau) \Leftrightarrow \tau \in T^{\mathcal{M}'} \Leftrightarrow [\![T(\tau)]\!] = t.$$

This completes case where φ is atomary. The cases where φ is non-atomary are easily proven using (2)-(4) (that is, (2) is used to prove the cases where $\varphi = \alpha \wedge \beta$ for some α, β, and so forth). ◊

Now we can prove that $\mathcal{M}'$ is a model of $U \cup A$ in the following way. First, since $\mathcal{M}'$ expands $\mathcal{M}$, it must be a model of U. Second, using the claim and (5) we get for all $T(\ulcorner\varphi\urcorner)$ with $\varphi \in M$,

$$\mathcal{M}' \models T(\ulcorner\varphi\urcorner) \Leftrightarrow [\![T(\ulcorner\varphi\urcorner)]\!] = t \Leftrightarrow G([\![\cdot]\!])(T(\ulcorner\varphi\urcorner)) = t \Leftrightarrow [\![\varphi]\!] = t \Leftrightarrow \mathcal{M}' \models \varphi$$

showing that all of the sentences in A are true in $\mathcal{M}'$. □

Proof of Lemma 3. Let G be any initial extension of $F_{\mathcal{L},M}$. We have to show that G has a total fixed point. For $f, g : \mathcal{L} \to \{t, f\}_\perp$ we define $f \subseteq g$ to mean that everywhere f is defined, g has the same value as f, that is, $\forall x(f(x) \neq \perp \to f(x) = g(x))$. It is easy to show that $F_{\mathcal{L},M}$, and therefore G, is monotone on $\mathcal{L} \to \{t, f\}_\perp$ with respect to the ordering $\subseteq$. Furthermore, $\mathcal{L} \to \{t, f\}_\perp$ is a ccpo[8] with respect to $\subseteq$, and therefore G must have a least fixed point $[\![\cdot]\!]$ (Gupta and Belnap 1993). To prove that $[\![\cdot]\!]$ is total, assume the opposite. Let D denote the dependence relation of G. This is of course identical to the dependence relation of $F_{\mathcal{L},M}$. Since D is conversely well-founded, the set of elements on which $[\![\cdot]\!]$ is undefined must have a D-maximal element φ. We must have $\varphi \notin \text{init}(F_{\mathcal{L},M})$ since otherwise—from the fact that G is an initial extension—we would get $[\![\varphi]\!] = G([\![\cdot]\!])(\varphi) \in \{t, f\}$. It is furthermore easily seen from the definition of $F_{\mathcal{L},M}$ that any non-initial element will have a non-empty dependence set. In particular, φ has a non-empty dependence set A. Since $[\![\varphi]\!] = \perp$, there must be some $\psi \in A$ s.t. $[\![\psi]\!] = \perp$. But this contradicts the D-maximality of φ since $\varphi D \psi$. □

[8] Chain complete partial order.

Proof of Lemma 5. First we prove that no set M' properly including M gives a conversely well-founded dependence relation for $F_{\mathcal{L},M'}$. Let thus M' be any such set. Let D denote the dependence relation of $F_{\mathcal{L},M'}$. Let D^* denote the transitive closure of D.[9] Since M' properly includes M, it must contain some sentence φ in which $T(x)$ is a subformula. That is, φ contains a subformula on the form $qx\psi(x)$ where $T(x)$ is a subformula of $\psi(x)$ and q is either $\exists$ or $\forall$. Since $qx\psi(x)$ is a subformula of φ there must furthermore be a sequence of formulas $\alpha_1, \ldots, \alpha_n$ such that

$$\varphi D \alpha_1 D \cdots D \alpha_n D q x \psi(x),$$

that is, $\varphi D^* q x \psi(x)$. From the clause (4) in the definition of $F_{\mathcal{L},M'}$ we see that

$$q x \psi(x) D \psi(\tau)$$

for all $\tau \in \text{Terms}(L)$. Since $T(x)$ is a subformula of $\psi(x)$ we must in addition have

$$\psi(\tau) D^* T(\tau)$$

for every $\tau \in \text{Terms}(L)$. Since $\varphi \in M'$, the clause (5) gives us

$$T(\ulcorner \varphi \urcorner) D \varphi.$$

Letting $\tau = \ulcorner \varphi \urcorner$ this means that

$$\varphi D^* q x \psi(x) D \psi(\ulcorner \varphi \urcorner) D^* T(\ulcorner \varphi \urcorner) D \varphi,$$

showing that D is not conversely well-founded (since $\varphi D^* \varphi$).

We now prove that the dependence relation D of $F_{\mathcal{L},M}$ is conversely well-founded. Assume the opposite, that is, assume there exists sentences $\varphi_1, \varphi_2, \varphi_3, \ldots$ such that

$$\varphi_1 D \varphi_2 D \varphi_3 \cdots \tag{9}$$

From the definition of $F_{\mathcal{L},M}$ we see that if $\alpha D \beta$ then one of the following three is the case

(i) $\alpha = \forall x \alpha'(x)$ for some formula α' and $\beta = \alpha'(\tau)$ for some term τ.

(ii) β is a subformula of α,

(iii) $\alpha = T(\ulcorner \beta \urcorner)$ and $\beta \in M$.

Claim (A). For all $i \geq 0$ there is a $j > i$ such that $\varphi_j = T(\ulcorner \varphi_{j+1} \urcorner)$.

Proof of claim. Assume the opposite. Then there is an i such that for all $j > i$, either $\varphi_j = \forall x \alpha(x)$ and $\varphi_{j+1} = \alpha(\tau)$ for some formula α and some term τ (corresponding to (i) above) or φ_{j+1} is a subformula of φ_j (corresponding to (ii) above). In both cases, φ_j will have higher syntactic complexity than φ_{j+1}. Thus $\varphi_{i+1}, \varphi_{i+2}, \ldots$ will be a chain of sentences of strictly decreasing syntactic complexity. This contradicts the chain (9) being infinite, and thus completes the proof. $\Diamond$

[9] The transitive closure R^* of a binary relation R is given by

$$x R^* x' \leftrightarrow \text{there exists a sequence of elements } x_1, \ldots, x_n \text{ such that } x R x_1 R x_2 R \cdots R x_n R x'.$$

Define a function $d : \mathrm{wff}(\mathcal{L}) \to \mathbb{N}$ by

$$d(\varphi) = \begin{cases} 1 + d(\psi) & \text{if } \varphi = T(\ulcorner \psi \urcorner) \text{ for some } \psi \\ 0 & \text{if } \varphi \text{ is any other atomary formula} \\ \max\{d(\psi) \mid \psi \text{ is a subformula of } \varphi\} & \text{otherwise.} \end{cases}$$

$d(\varphi)$ is called the T**-degree** of φ. The well-definedness of d given by these recursive clauses is ensured by the fact that for all ψ, $\ulcorner T(\ulcorner \psi \urcorner) \urcorner > \ulcorner \psi \urcorner$. The function d can thus be defined by recursion on the Gödel number of φ.

Now choose $j > 0$ such that φ_j satisfies claim (A), that is, $\varphi_j = T(\ulcorner \varphi_{j+1} \urcorner)$. We then have the following claim

Claim (B). The T-degree is monotonically decreasing in the chain $\varphi_{j+1} D \varphi_{j+2} D \varphi_{j+3} \cdots$.

Proof of claim. Assume there is a φ_i such that φ_{i+1} has greater T-degree than φ_i. Then $\varphi_i = \forall x \alpha(x)$ and $\varphi_{i+1} = \alpha(\tau)$ for some formula α and term τ (corresponding to (i) above. In both of the cases (ii) and (iii) the T-degree is constant or decreasing). But $\alpha(\tau)$ can only have higher T-degree than $\forall x \alpha(x)$ if $T(x)$ is a subformula of $\alpha(x)$ (otherwise instantiating x with τ will leave the T-degree unchanged). Now let k be the greatest number less than i for which $\varphi_{k-1} = T(\ulcorner \varphi_k \urcorner)$ with $\varphi_k \in M$. The choice of j guarantees us the existence of such a number. Then φ_i must be a subformula of φ_k, and therefore $T(x)$ must be a subformula of φ_k as well. But this contradicts φ_k being in M, which concludes the proof. ◊

We now get a contradiction from claim (B): since the T-degree is monotonically decreasing in the chain $\varphi_{i+1} D \varphi_{i+2} D \varphi_{i+3} \cdots$, the T-degree must be constant from some sentence on. But this immediately contradicts claim (A). □

Proof of Lemma 9. Assume M contains a sentence

$$\psi = \forall x \left(A(x) \vee \neg T(x) \right).$$

where $A \neq T$. We need to show that the set of sentences

$$P = \{ \varphi \leftrightarrow T(\ulcorner \varphi \urcorner) \mid \varphi \in M \}$$

is not universally. To prove this, let U be the theory containing the axiom

$$\forall y \left(y = \ulcorner \psi \urcorner \leftrightarrow \neg A(y) \right)$$

and axioms $\vdash \bar{n} \neq \bar{m}$ for all $n \neq m$. U is obviously consistent. We now only have to show that $U \cup P$ is not consistent, which proves P not to be universally consistent. To obtain a contradiction, assume that $U \cup P$ is consistent. Then it has a model $\mathcal{M}$. This gives us the following sequence of implications

$$\mathcal{M} \models \psi \Rightarrow \mathcal{M} \models \forall x \left(A(x) \vee \neg T(x) \right) \Rightarrow \mathcal{M} \models \forall x \left(x \neq \ulcorner \psi \urcorner \vee \neg T(x) \right) \Rightarrow$$
$$\mathcal{M} \models \ulcorner \psi \urcorner \neq \ulcorner \psi \urcorner \vee \neg T(\ulcorner \psi \urcorner) \Rightarrow \mathcal{M} \models \neg T(\ulcorner \psi \urcorner) \Rightarrow \mathcal{M} \models \neg \psi,$$

which shows that $\mathcal{M} \not\models \psi$. At the same time we have

$$\mathcal{M} \models \neg \psi \Rightarrow \mathcal{M} \models \exists x \left(\neg A(x) \wedge T(x) \right) \Rightarrow \mathcal{M} \models \exists x \left(x = \ulcorner \psi \urcorner \wedge T(x) \right) \Rightarrow$$
$$\mathcal{M} \models \ulcorner \psi \urcorner = \ulcorner \psi \urcorner \wedge T(\ulcorner \psi \urcorner) \Rightarrow \mathcal{M} \models T(\ulcorner \psi \urcorner) \Rightarrow \mathcal{M} \models \psi,$$

which shows that $\mathcal{M} \not\models \neg\psi$. Now we have both $\mathcal{M} \not\models \psi$ and $\mathcal{M} \not\models \neg\psi$, which is a contradiction.

□

Thomas Bolander
Informatics and Mathematical Modelling
Technical University of Denmark
Richard Petersens Plads, Building 321
DK-2800 Lyngby, Copenhagen, Denmark
tb@imm.dtu.dk

References

Bealer, G. 1982. *Quality and Concept*. Oxford University Press.

des Rivières, J. and Levesque, H. J. 1986. The consistency of syntactical treatments of knowledge. In *Theoretical aspects of reasoning about knowledge*, pages 115–130. Morgan Kaufmann.

Feferman, S. 1984. Toward useful type-free theories I. *The Journal of Symbolic Logic*, 49(1):75–111.

Gupta, A. and Belnap, N. 1993. *The Revision Theory of Truth*. MIT Press.

McGee, V. 1992. Maximal consistent sets of instances of Tarski's schema (T). *Journal of Philosophical Logic*, 21(3):235–241.

Mendelson, E. 1997. *Introduction to Mathematical Logic*. Chapman & Hall, 4 edition.

Perlis, D. 1985. Languages with self-reference I. *Artificial Intelligence*, 25:301–322.

Stoltenberg-Hansen, V., Lindström, I., and Griffor, E. R. 1994. *Mathematical theory of domains*. Cambridge University Press, Cambridge.

Tarski, A. 1956. The concept of truth in formalized languages. In *Logic, semantics, metamathematics*. Hackett Publishing Co., Indianapolis, IN. Papers from 1923 to 1938.