Truth II[1]

Marco Degano

Philosophical Logic 2024
12 November 2024

[1] Slides also based on teaching material by Frank Veltman

Readings

Optional:

- Tarski, Alfred (1944). The semantic conception of truth: and the foundations of semantics. *Philosophy and phenomenological research*, 4(3), 341-376.
- Kripke, Saul (1975). Outline of a theory of truth. *The journal of philosophy*, 72(19), 690-716.

Plan

1. Tarski's Theory of Truth

2. Kripke's Theory of Truth

The Liar Paradox

T-in: $\models \phi \to T(\ulcorner \phi \urcorner)$

T-out: $\models T(\ulcorner \phi \urcorner) \to \phi$

Liar sentence: $\phi = \neg T(\ulcorner \phi \urcorner)$

1. $T(\ulcorner \phi \urcorner) \vee \neg T(\ulcorner \phi \urcorner)$	LEM
2. $T(\ulcorner \phi \urcorner)$	Hyp
3. ϕ	T-out, 2
4. $\neg T(\ulcorner \phi \urcorner)$	Meaning of ϕ, 3
5. $T(\ulcorner \phi \urcorner) \wedge \neg T(\ulcorner \phi \urcorner)$	2, 5 $\wedge$-I
6. $\neg T(\ulcorner \phi \urcorner)$	Hyp
7. ϕ	Meaning of ϕ, 6
8. $T(\ulcorner \phi \urcorner)$	T-in, 7
9. $T(\ulcorner \phi \urcorner) \wedge \neg T(\ulcorner \phi \urcorner)$	6, 8 $\wedge$-I
10. $T(\ulcorner \phi \urcorner) \wedge \neg T(\ulcorner \phi \urcorner)$	Reasoning by cases
11. $\bot$	explosion, 10

T-schema

For Tarski, any adequate theory of truth must satisfy the T-schema or Convention T:

for any $\phi \in \mathcal{L}$, we have $T(\ulcorner\phi\urcorner) \leftrightarrow \phi$

The Liar paradox indicates that accepting the T-schema and self-reference leads to an inconsistent (or trivial) theory.

Tarski's Undefinability Theorem

The Liar Paradox relates to Tarski's Undefinability Theorem.[2]

Given the language of arithmetic augmented with a truth predicate, the following result follows from a standard application of the Diagonalization Lemma:[3]

Let T be any consistent theory that contains **PA**. Then, the property of being a member of T is not representable in T.

[2] See Schlöder (2020) for an accessible proof of the theorem.

[3] Interestingly, different paradoxes can be given a uniform representation exploiting this technique. See Yanofsky, N. S. (2003). *A Universal Approach to Self-Referential Paradoxes, Incompleteness, and Fixed Points.*

Outline

The paradox and our assumptions

The Liar paradox rests on three fundamental assumptions:

1. $\mathcal{L}$ is **semantically closed**: names for sentences and predicates that apply to such sentences.

2. Classical logic holds in $\mathcal{L}$

3. Self-reference is allowed in $\mathcal{L}$.

Tarski wants to preserve (2) and argues that (3) is not sufficient (recall that the Liar Paradox can be generated without self-reference).

Tarski targets the first point.

Truth of the object language

We fix a language $\mathcal{L}$. All sentences in $\mathcal{L}$ have a meaning (they are interpreted as true or false). Example: 2+2 = 4

A meta-language should be able to talk about the sentences in $\mathcal{L}$. Example: '2+2=4' is true.

But what about the sentence ' '2+2=4' is true' is true?

$\mathcal{L}_0$: language with no truth predicate

$\mathcal{L}_1$: metalanguage of $\mathcal{L}_0$: a truth predicate over $\mathcal{L}_0$

$\mathcal{L}_2$: ...

⋮

Liar Paradox is blocked!

Hierarchy of languages

$\mathcal{L}_0$ interpreted by a classical model $M_0 = \langle D, I \rangle$

We augment $\mathcal{L}_0$ with a truth predicate T_0 for the sentences $\phi \in \mathcal{L}_0$

$\mathcal{L}_1$ interpreted by a model $M_1 = \langle M_0, \mathcal{T}_0 \rangle$, with $\mathcal{T}_0$ being the set of true sentences in $\mathcal{L}_0$.

$M_1 \models T_0('\phi') \text{ iff } '\phi' \in \mathcal{T}_0 \text{ iff } M_0 \models \phi$

Language	Model	T-schema
⋮	⋮	⋮
$\mathcal{L}_{i+1}$	$M_{i+1} = \langle M_i, \mathcal{T}_i \rangle$	$T_i('\phi') \leftrightarrow \phi$ for $\phi \in \mathcal{L}_i$
⋮	⋮	⋮
$\mathcal{L}_2$	$M_2 = \langle M_1, \mathcal{T}_1 \rangle$	$T_1('\phi') \leftrightarrow \phi$ for $\phi \in \mathcal{L}_1$
$\mathcal{L}_1$	$M_1 = \langle M_0, \mathcal{T}_0 \rangle$	$T_0('\phi') \leftrightarrow \phi$ for $\phi \in \mathcal{L}_0$
$\mathcal{L}_0$	$M_0 = \langle D, I \rangle$	–

[Lui 2019]

Critical Remarks: Multiplicity of Truths

Hierarchy of Languages $\longrightarrow$ Hierarchy of Truths

But there seems to be only one conception of truth in natural languages.

Tarski recognized this and claimed that natural languages are distinct in this respect with formal languages.

[Sher 1999]

Critical Remarks: Kripke's criticism

(j) John: 'Most (more than half) of Nixon's utterances about Watergate are false.'

(n) Nixon: 'Everything John said about Watergate is true.'

Except (j), all John's utterances about Watergate are true.

Except (n), half of Nixon's utterances about Watergate are true, and half of them are false.

Then (j) is true iff (j) is false; and (n) is true iff (n) is false.

But (j) has to be one level higher than all of Nixon's utterances, AND (n) has to be one level higher that all of John's utterances.

Outline

Kripke: informal picture

We begin with a language that does not **initially** interpret the truth predicate.

We **extend the interpretation** of the truth predicate by adding more sentences that are determined to be true.

This process of adding true sentences is **monotonic** - once a sentence is deemed true, it remains true.

A **fixed point** is a point at which the interpretation of the truth predicate stabilizes and no longer changes with further extensions.

The Liar sentence is treated as **ungrounded** in the fixed point: it neither comes out true nor false.

The Language

Let $\mathcal{L}$ be a first order language with

- a one-place predicate T. $T(x)$ for x *is true*.
- for every sentence ϕ, an individual term $'\phi'$ standing for the name of the sentence ϕ.

We write $S_{\mathcal{L}}$ for the set of all sentences of $\mathcal{L}$.

Example

How to capture the liar sentence in this language?

An individual constant l standing for the name of $\neg T(l)$

Try to formalize 'Everything that John said is false' in the language.

Semantics

A model M for $\mathcal{L}$ is a triple $\langle D, I, \mathcal{T} \rangle$ such that

- $S_{\mathcal{L}} \subseteq D$;
- I is a function assigning:
  - an element $I(a)$ of D to all individual constants a
  - the sentence ϕ to each term $'\phi'$
  - and a total function from D^n in $\{0, 1\}$ to every n-ary predicate except T.

$\mathcal{T} \subseteq S_{\mathcal{L}} \times \{0, 1\}$

$\mathcal{T}$ will serve as the interpretation for the truth predicate T.

Semantic Clauses

$M \vDash P(a_0 \ldots a_n)$ iff $I(P)\left(\left\langle I\left(a_0\right), \ldots, I\left(a_n\right)\right\rangle\right) = 1$
$M \nvDash P(a_0 \ldots a_n)$ iff $I(P)\left(\left\langle I\left(a_0\right), \ldots, I\left(a_n\right)\right\rangle\right) = 0$

$M \vDash \neg\phi$ iff $M \nvDash \phi$
$M \nvDash \neg\phi$ iff $M \vDash \phi$

$M \vDash \phi \wedge \psi$ iff $M \vDash \phi$ and $M \vDash \psi$
$M \nvDash \phi \wedge \psi$ iff $M \vDash \phi$ or $M \vDash \psi$

$M \vDash \phi \vee \psi$ iff $M \vDash \phi$ or $M \vDash \psi$
$M \nvDash \phi \vee \psi$ iff $M \nvDash \phi$ and $M \nvDash \psi$

$M \vDash \phi \rightarrow \psi$ iff $M \nvDash \phi$ or $M \vDash \psi$
$M \nvDash \phi \rightarrow \psi$ iff $M \vDash \phi$ and $M \nvDash \psi$

Semantic Clauses

$M \vDash \exists x\phi$ iff $M \vDash [a/x]\phi$ for some individual constant a
$M \nvDash \exists x\phi$ iff $M \nvDash [a/x]\phi$ for all individual constants a

$M \vDash \forall x\phi$ iff $M \vDash [a/x]\phi$ for all individual constants a
$M \nvDash \forall x\phi$ iff $M \nvDash [a/x]\phi$ for some individual constant a.

And **most importantly**:

$M \vDash T(a)$ iff $\langle I(a), 1\rangle \in \mathcal{T}$
$M \nvDash T(a)$ iff $\langle I(a), 0\rangle \in \mathcal{T}$

We will be requiring that for each a s.t. $I(a) = \phi$ for some ϕ, we cannot have that both $\langle I(a), 1\rangle \in \mathcal{T}$ and $\langle I(a), 0\rangle \in \mathcal{T}$.

Transparent Truth

Can we have the following?

$M \vDash T\,('\phi') \text{ iff } M \vDash \phi$

No, if we allow models with a liar sentence and if $\mathcal{T}$ is a total function.

The Liar sentence

Consider l in models such that $I(l) = \neg T(l)$.

Then we have:

$M \vDash T\left('\neg T(l)'\right)$ iff $\langle I\left('\neg T(l)'\right), 1\rangle \in \mathcal{T}$, iff $\langle \neg T(l), 1\rangle \in \mathcal{T}$,

whereas

$M \vDash \neg T(l)$ iff $M \nvDash T(l)$ iff $\langle I(l), 0\rangle \in \mathcal{T}$ iff $\langle \neg T(l), 0\rangle \in \mathcal{T}$

Revaluation

Let $M = \langle D, I, \mathcal{T} \rangle$ be a model.

The revaluation of M is the model $J(M) = \langle D, I, J(\mathcal{T}) \rangle$ such that

$\langle \phi, 1 \rangle \in J(\mathcal{T})$ iff $M \vDash \phi$;

$\langle \phi, 0 \rangle \in J(\mathcal{T})$ iff $M \nvDash \phi$.

$J(\mathcal{T})$ is called a revaluation of $\mathcal{T}$.

$\mathcal{T}$ is coherent iff $\mathcal{T} \subseteq J(\mathcal{T})$.

Example

Consider $M = \langle D, I, \mathcal{T} \rangle$ with $\mathcal{T} = \varnothing$. Let P be a predicate different from T, and assume that $I(a) \in I(P)$.

It is easy to check that

- $M \vDash P(a)$ $\quad$ $M \not\vDash T('P(a)')$
- $J(M) \vDash P(a)$ $\quad$ $J(M) \vDash T('P(a)')$
  $J(M) \not\vDash T('T('P(a)')')$
- $J(J(M)) \vDash P(a)$ $\quad$ $J(J(M)) \vDash T('P(a)')$
  $J(J(M)) \vDash T('T('P(a)')')$

Example: the Liar

'This sentence is false'

Consider $M = \langle D, I, \mathcal{T} \rangle$ such that $I(l) = \neg T(l)$ and $\mathcal{T} = \{\langle \neg T(l), 0 \rangle\}$.

Notice that $\mathcal{T}$ is incoherent:

$\langle \neg T(l), 0 \rangle \in J(\mathcal{T})$ iff $M \nvDash \neg T(l)$ iff $M \vDash T(l)$ iff $\langle I(l), 1 \rangle \in \mathcal{T}$ iff $\langle \neg T(l), 1 \rangle \in \mathcal{T}$

Note that for this $I, \mathcal{T} = \{\langle \neg T(l), 1 \rangle\}$ is incoherent, too.

The Truth-teller

'This sentence is true'

Consider $M = \langle D, I, \mathcal{T} \rangle$ such that $I(t) = T(t)$ and $\mathcal{T} = \{\langle T(t), 1 \rangle\}$.

$\mathcal{T}$ is coherent.

$\mathcal{T} = \{\langle T(t), 0 \rangle\}$ is coherent, too.

Stability of Revaluation

Let $M = \langle D, I, \mathcal{T} \rangle$ and $M' = \langle D, I, \mathcal{T}' \rangle$ be two models such that $\mathcal{T} \subseteq \mathcal{T}'$.

Prove that $J(\mathcal{T}) \subseteq J(\mathcal{T}')$.

Revaluation Sequence

Let $M = \langle D, I, \mathcal{T} \rangle$ be a model.

By transfinite induction we define the Kripkean revaluation sequence $M_0 = \langle D, I, \mathcal{T}_0 \rangle, M_1 = \langle D, I, \mathcal{T}_1 \rangle, \ldots, M_\omega = \langle D, I, \mathcal{T}_\omega \rangle, M_{\omega+1} = \langle D, I, \mathcal{T}_{\omega+1} \rangle, \ldots$, as follows

$\mathcal{T}_0 = \mathcal{T}$

$\mathcal{T}_{\sigma+1} = J(\mathcal{T}_\sigma)$

If σ is a limit ordinal, then $\mathcal{T}_\sigma = \bigcup_{\tau<\sigma} \mathcal{T}_\tau$

Monotonicity

Let $M = \langle D, I, \mathcal{T} \rangle$ be a model with coherent $\mathcal{T}$. Then

$$\sigma < \tau \Rightarrow \mathcal{T}_\sigma \subseteq \mathcal{T}_\tau$$

Fixed Points

Let $M = \langle D, I, \mathcal{T} \rangle$ be a model with coherent $\mathcal{T}$. There exists a unique ordinal number ρ such that

- for all $\sigma < \tau \leq \rho, \mathcal{T}_\sigma \subsetneq \mathcal{T}_\tau$
- for all $\sigma \geq \rho, \mathcal{T}_\sigma = \mathcal{T}_\rho$

Note that $J(\mathcal{T}_\rho) = \mathcal{T}_\rho$, which is why $\mathcal{T}_\rho$ is called the fixed point of J generated by $\mathcal{T}$. We will often write $\mathcal{T}^*$ for this point.

Minimal Fixed Point

Let $\mathcal{T}$ and $\mathcal{T}'$ be coherent and $\mathcal{T} \subseteq \mathcal{T}'$

Prove that $\mathcal{T}^* \subseteq \mathcal{T}'^*$

Thus $\varnothing^*$ is the minimal fixed point.

Kinds of sentences

fixed points behaviour		**example**
true in all, false in no	**grounded** true	
true in some, false in no		(Tutorial)[4]
false in all, true in no	**grounded** false	
false in some, true in no		(Assignment)
true in no, false in no	**paradoxical**	the liar
true in some, false in some	**biconsistent**	the truth teller

[4] $\phi = T(t) \vee \neg T(t)$ in models in which $I(t) = T(t)$.

Revenge

This sentence is false or neither true nor false.

Suppose we were to add a new one place sentential operator to the language with the following semantics:

$M \models\sim \phi$ iff $M \not\models \phi$

$M \nvDash\sim \phi$ iff $M \models \phi$

Now, consider $l^{\sim}$ such that $I\left(l^{\sim}\right) =\sim T(l^{\sim})$.

Then we have: $M \models T('\sim T(l^{\sim})')$ iff $\langle I\left('\sim T(l^{\sim})'\right), 1\rangle \in \mathcal{T}$, iff $\langle \sim T\left(l^{\sim}\right), 1\rangle \in \mathcal{T}$

whereas $M \models\sim T\left(l^{\sim}\right)$ iff $M \not\models T\left(l^{\sim}\right)$ iff $\langle I\left(l^{\sim}\right), 1\rangle \notin \mathcal{T}$ iff $\langle \sim T\left(l^{\sim}\right), 1\rangle \notin \mathcal{T}$