PHILOSOPHY OF LOGIC AND LANGUAGE

WEEK 5: TARSKI ON TRUTH

JONNY MCINTOSH

INTRODUCTION

For the second half of term, I want to look at issues in the philosophy of logic and language arising out of the work of Alfred Tarski.

Alfred Tarski (1901-1983)

Tarski was a Polish (and later American) mathematician, logician, and philosopher.

His work in the 1930s laid the foundations for the modern conception of logic.

The plan for the next four weeks:

5. Tarski on truth
6. The Liar Paradox
7. Logical consequence
8. Logical constants

BACKGROUND

The notion of truth seems to lie at the heart of various notions that are central to thinking about logic and mathematics, e.g. validity, consistency, completeness.

But the notion of truth seems to involve us in contradiction, giving rise to so-called **SEMANTIC PARADOXES** such as The Liar Paradox.

Here's a brief illustration. Let λ be the sentence 'λ is not true'. Then:

1. 'λ is not true' is true IFF λ is not true
2. λ = 'λ is not true'
3. So, λ is true IFF λ is not true

Tarski was worried that, unless such paradoxes can be resolved, metatheoretical results invoking the notion of truth or other notions that depend on it will remain suspect.

But how can it be resolved? The second premise is undeniable, and the first premise is an instance of a schema that seems central to the concept of truth: *S is true IFF P*.

(This is a **SCHEMA**, and not a sentence. We obtain sentences *from* it by replacing 'S' with the name of an English sentence and 'P' with that sentence.)

Tarski thinks the problem arises because languages like English are **SEMANTICALLY CLOSED**: for each sentence S, they contain another sentence S’ that attributes truth (untruth) to S.

But we need to be able to talk about the true sentences of a languages if we are to even formulate metatheoretical claims like completeness.

Tarski's solution: talk about the true (untrue) sentences of one language, the **OBJECT LANGUAGE**, in different language, the **METALANGUAGE**.

How? Define a predicate true_L in the metalanguage that applies to all and only the true sentences of the object language.

THE PROBLEM

I'll talk you through the various requirements that such a definition has to meet, in Tarski's view.

FORMAL CORRECTNESS

First, a satisfactory definition of a predicate, true_L, must be **FORMALLY CORRECT**. In other words:

It must be provably equivalent to a sentence of the form:

- $\forall x$, $\mathrm{true}_L(x)$ IFF $\phi(x)$,

where ϕ doesn't contain true_L or any related predicate.

(Otherwise put: a satisfactory definition of true_L must be provably equivalent to an **EXPLICIT** definition.)

This requirement ensures that the definition is non-circular, but it doesn't ensure that the result is a *truth* predicate.

To see this, consider the definition:

- $\forall x$, $\mathrm{true}_L(x)$ IFF $x = x$.

This is formally correct, but the predicate defined applies to *every* sentence of L, not just the true ones.

MATERIAL ADEQUACY

So we need a definition which is not just formally correct, but also **MATERIALLY ADEQUATE**, i.e. applies to all and only the true sentences of L.

Tarski suggests that a (formally correct) definition of $\textit{true}_L$ for a given language L is materially adequate IFF...

...it entails, for each sentence of L, a relevant instance of **SCHEMA T**:

- S is true_L IFF P,

where the relevant instances are obtained by replacing 'S' with a name of the sentence and 'P' by a translation of the sentence in the metalanguage, M.

Tarski's suggestion here is known as **CONVENTION T**. A better name might be **CRITERION T**. Be careful not confuse it with Schema T!

What is the idea behind Convention T? Consider an analogy. Suppose that teenagers have started using a new word — 'peng'.

The Chicken Connoisseur

You're trying to work out what on earth 'peng' means. Could it possibly be a truth predicate? Under what conditions could we determine that it is?

It seems that it can *only* be a truth predicate for English if every relevant instance of **SCHEMA P** is true:

- S is peng IFF P

As before, the relevant instances are obtained by replacing 'S' with a name of the sentence and 'P' by a translation of the sentence in the metalanguage, M.

For example, it will have to be the case that 'Snow is white' is peng IFF snow is white, and that 'Grass is blue' is peng IFF grass is blue.

This seems to show that a *necessary* condition on 'peng' being a truth predicate for English that each relevant instance of Schema P is true.

On the other hand, it also seems to be a sufficient condition. Why? Well, if every relevant instance is true then, if a sentence says that P, it is peng IFF P.

But if a sentence says that P, it is *true* IFF P. So *if* every relevant sentence of Schema P is true then: a sentence of English is peng IFF it is true.

TARSKI'S HIERARCHY

Various constraints on the relationship between L and M fall out of Convention T.

FIRST, M must contain the resources required for *referring* to each sentence of L.

SECOND, M has to include the sentences of L, or at least *translations* of them.

But this doesn't in and of itself rule out the possibility that M is identical to L. So why does Tarski think that it can't be?

Suppose that M is identical to L, and meets the two constraints I just mentioned. Then the predicate $true_L$ can be defined within L itself.

So long as L contains a negation operator, ¬, it will also contain a Liar sentence: a sentence of the form $\neg true_L(\lambda)$ whose name in L is of the form λ.

And by Convention T, the definition of $true_L$ entails a theorem of the form, $\neg true_L(\lambda)$ *IFF* $true_L(\lambda)$.

But now we have the two premises we need for the Liar Paradox. So as long as we've got a classical logic, a contradiction follows.

So, if we want to (a) keep classical logic and (b) allow that L contains ¬, we need to give up the assumption that M is identical to L.

Similarly, M has to be distinct from the meta-metalanguage in which we define *its* truth predicate. We need an entire *hierarchy* of languages.

The truth predicate we define in the metalanguage only applies to sentences of the object language, and so to sentences that don't contain that predicate.

TARSKI'S SOLUTIONS

Tarski doesn't just tell us what a definition of $\textit{true}_L$ will have to look like. He shows how to construct it for various given languages L.

FINITE LANGUAGES

It's very easy in some simple cases. Suppose L_1 contains just two sentences, '1 + 1 = 2' and '1 + 1 = 3'.

Assuming these have their ordinary meanings, then, using English as our metalanguage, our definition might be:

∀x (x is true_{L_1} IFF

- x = '1 + 1 = 2' and one plus one is two OR
- x = '1 + 1 = 3' and one plus one is three)

Questions to consider:

- Is this formally correct?
- If so, is it materially adequate?

RECURSIVE DEFINITIONS

Defining $\textit{true}_{L_1}$ is straightforward. We just list out the truth conditions of the sentences of L_1.

But now consider the language, L_2, that extends L_1 with the connectives, ¬ and ∧.

These operators can be iterated, so L_2 contains infinitely many sentences.

In this case, Tarski suggests we give a *recursive* definition:

- We first define $\textit{true}_{L_2}$ for atomic sentences
- We then define $\textit{true}_{L_2}$ for complex sentences

A recursive definition of *true*$_{L_2}$ might be:

∀x, y, z (x is true$_{L_2}$ IFF

- x = '1 + 1 = 2' and one plus one is two OR
- x = '1 + 1 = 3' and one plus one is three OR
- y is a sentence and x = ⌜¬y⌝ is not true$_{L_2}$ OR
- y and z are sentences and x = ⌜(y∧z)⌝ and x and y are true$_{L_2}$)

Is this formally correct? Well, it's not *itself* an explicit definition. But it is provably equivalent to one.

This is a due to a result from Frege and Dedekind: given some mathematical machinery, certain recursive definitions can be turned into explicit ones.

Something to think about: how do we know this is materially adequate?

QUANTIFICATION

We can't specify the truth conditions of '∃x (x + x = 2)' in terms of the truth values of its parts: its parts don't *have* truth values.

Tarski offers two methods for dealing with this case. The first involves replacing variables with names of the objects in the domain of quantification.

But this won’t always work, as in general some objects in the domain of quantification won’t have names. (Suppose the domain is $\mathbb{R}$.)

So in the general case, Tarski's method involves the notion of a **VARIABLE ASSIGNMENT**, and treats the variables themselves as a kind of temporary name.

Something to think about: does this differ from Frege's treatment in any essential respect?

PHILOSOPHICAL SIGNIFICANCE

Tarski shows us how to give explicit definitions of predicates that are co-extensive with various substitution instances of the predicate 'is a true sentence of L'.

This is an impressive achievement, and enables us to talk about the true sentences of various languages without risking paradox. Does he also do more?

DEFINITION AND EXPLICATION

Does Tarski provide the means for defining the concept of truth, i.e. for defining predicates that have the same meaning as the English predicate 'is true'?

An initial problem is that the English predicate 'is true' appears to apply to propositions, not sentences.

So ask a different question. Does Tarski provide the means for defining predicates that have the same meaning as the English predicate 'expresses a truth'?

It seems not. The English predicate 'expresses a truth' applies to the sentences of a *range* of languages, including English!

Tarski’s predicates only apply to one language, and do not apply to sentences of the language to which they belong.

If Tarski's diagnosis of the Liar Paradox is right, this means his predicates *cannot* be co-extensive with 'is true' or 'expresses a truth'.

Perhaps we could more modestly take Tarski as providing the means for **EXPLICATING** the concept of truth.

To explicate a concept is to provide a predicate that can replace predicates expressing that concept in legitimate theoretical contexts and lacks their defects.

PROBLEMS

But various problems might be raised. First, there is **THE PROBLEM OF TRUTH-CONDITIONAL SEMANTICS**.

One of the main approaches to meaning in philosophy of language and linguistics is that of truth-conditional semantics.

On this approach, roughly, a definition of truth for a language can be used to tell us what the sentences of that language mean.

And the problem is that it seems that a Tarskian definition of 'true$_L$' cannot be used to do that.

Why not? For a definition of 'true$_L$' to tell us what the sentences of L mean, we need to *already* know what 'true$_L$' means.

(Imagine trying to use a definition of 'peng' to tell someone what the sentences of French mean. Unless they know that 'peng' means 'true', they'll be mystified.)

But telling us what 'true$_L$' means is the task of the Tarskian definition. So it seems Tarskian definitions can't be used in truth-conditional semantics.

More exactly: a Tarskian definition cannot *both* (a) tell us what the sentences of the object language mean *and* (b) tell us what the predicate being defined means.

Second, there's **THE PROBLEM OF MODAL DIFFERENCE**. Recall the definition of 'true$_{L_1}$':

∀x (x is true$_{L_1}$ IFF

- x = '1 + 1 = 2' and one plus one is two OR
- x = '1 + 1 = 3' and one plus one is three)

If we replace 'true$_{L_1}$' in '1 + 1 = 2' is true$_{L_1}$ with its definiens, we get:

('1 + 1 = 2' is '1 + 1 = 2' and one plus one is two) OR ('1 + 1 = 2' is '1 + 1 = 3 and one plus one is three)

But now, the sentence '1 + 1 = 2' has *this* (disjunctive) property in all worlds in which one plus one is two — i.e. *all* worlds.

But surely '1 + 1 = 2' is not *true* in all worlds. For example, isn't it false in worlds in which '2' refers to three, and all the other words have their actual meaning?

Third, there's **THE PROBLEM OF NON-PROJECTABILITY**. Tarski's definitions don't *project*.

They don't tell us under what conditions other truth predicates, for other languages, hold of the sentences of those languages.

Fourth, there is **THE PROBLEM OF EPISTEMIC DIFFERENCE**. Knowing a sentence's truth conditions gives us at least *negative* information about its meaning.

For example, knowing that 'La neige est blanche' is true IFF snow is white tells us that it *doesn't* mean the same as 'Snow is not white'.

But knowing the conditions under which, say, '1 + 1 = 2' is true$_L$ doesn’t even seem to provide that.

Compare: knowing that '1 + 1 = 2' is peng IFF one plus one is two seems perfectly compatible with it meaning the same as 'one plus one is not two'.