

Formal Theories of Truth I
THE THEORY OF SYNTAX AND
DIAGONALISATION

Volker Halbach

17th February 2019

There is something fishy about the liar paradox:

(1) (1) is not true.

Somehow the sentence 'says' something about itself, and when people are confronted with the paradox for the first time, they usually think that this feature is the source of the paradox.

Self-reference

However, there are many self-referential sentences that are completely unproblematic:

(2) (2) contains 5 occurrences of the letter 'c'.

If (1) is illegitimate because of its self-referentiality, then (2) must be illegitimate as well. Moreover, the effect that is achieved via the label '(1)' can be achieved without this device. At the same time one can dispense with demonstratives like 'this' that might be used to formulate the liar sentence:

This sentence is not true.

In fact, the effect can be achieved using weak arithmetical axioms only. And the axioms employed are beyond any (serious) doubt. This was shown by Gödel.

Arithmetic

The approach via arithmetic is indirect. Arithmetic talks about numbers, not about sentences. Coding sentences and expressions by numbers allows to talk about the numerical codes of sentences and therefore arithmetic is indirectly about sentences.

My approach here avoids this detour via numbers. I present a theory of expressions that is given by some (as I hope) obvious axioms on expressions. The trick (diagonalization) that is then used for obtaining a self-referential sentence is the same as in the case of arithmetic.

The alphabet

In the following I describe a language \mathcal{L} . An expression of \mathcal{L} is an arbitrary finite string of the following symbols. Such strings are also called expressions of \mathcal{L} .

Definition

The symbols of \mathcal{L} are:

- 1 infinitely many variable symbols v, v_1, v_2, v_3, \dots
- 2 predicate symbols $=$ and T ,
- 3 function symbols q, \wedge and sub ,
- 4 the connectives \neg, \rightarrow and the quantifier symbol \forall ,
- 5 auxiliary symbols $($ and $)$,
- 6 possibly finitely many further function and predicate symbols, and
- 7 If e is a string of symbols then \bar{e} is also a symbol. \bar{e} is called a quotation constant.

All the mentioned symbols are pairwise different.

Notational conventions

In the following I shall use x , y and z as (meta-)variables for variables. Thus x may stand for any symbol v, v_1, v_2, \dots . It is also assumed that x, y etc stand for different variables. Moreover, it is always presupposed variable clashes are avoided by renaming variables in a suitable way.

It is important that \bar{a} is a single symbol and not a string of more than one symbols even if a itself is a string built from several symbols.

A string of symbols of \mathcal{L} is any string of the above symbols. Usually I suppress mention of \mathcal{L} . The empty string is also a string.

We shall now define the notions of a term and of a formula of \mathcal{L} .

Definition

The \mathcal{L} -terms are defined as follows:

- 1 All variables are terms.
- 2 If e is a string of symbols, then \bar{e} is a term.
- 3 If t, r and s are terms, then $q(t), (s \wedge t), \text{sub}(r, s, t)$ are terms, and similarly for all further function symbols

The empty string

Since the empty string is a string of symbols $\bar{\quad}$ is a term. Since $\bar{\quad}$ looks so odd, I shall write $\underline{0}$ for $\bar{\quad}$. From an ontologically point of view the empty string is a weird thing. One might be inclined to say that it is not anything. I have only a pragmatic excuse for assuming the empty string: it is useful, though not indispensable.

What the empty string is for the expressions is the number zero for the natural numbers. It is not hard to see that 0 is useful in number theory.

Formulæ, sentences, free and bound occurrences of variables are defined in the usual way.

Example

- 1 $\forall v_3(v_3 = \overline{\wedge \forall} \wedge T\overline{v_3})$ is a sentence.
- 2 $\overline{v_{12}} = \overline{\neg T \neg}$ is a sentence, i.e., the formula does not feature a free variable.

A theory of expressions

The theory \mathcal{A} which will be described in this section is designed in order to obtain smooth proofs. I have not aimed at a particularly elegant axiomatization.

A simple intended model of the theory has all expressions of \mathcal{L} as its domain. The intended interpretation of the function symbols will become clear from the axioms A1–A4 except for the interpretation of sub . I shall return to sub below.

The axioms

All instances of the following schemata and rules are axioms of the theory \mathcal{A} :

Definition

- Ⓐ₁ all axioms and rules of first-order predicate logic including the identity axioms.
- Ⓐ₂ $\overline{a} \overline{b} = \overline{ab}$, where a and b are arbitrary strings of symbols.
- Ⓐ₃ $q(\overline{a}) = \overline{\overline{a}}$
- Ⓐ₄ $\text{sub}(\overline{a}, \overline{b}, \overline{c}) = \overline{d}$, where a and c are arbitrary strings of symbols, b is a single symbol (or, equivalently, a string of symbols of length 1), and d is the string of symbols obtained from a by replacing all occurrences of the symbol b by the strings c .

Probably we won't need the following axioms:

Definition (additional axioms)

- A5 $\forall x \forall y \forall z ((x \wedge y) \wedge z) = (x \wedge (y \wedge z))$
- A6 $\forall x \forall y (x \wedge y = \underline{0} \rightarrow x = \underline{0} \wedge y = \underline{0})$
- A7 $\forall x \forall y (x \wedge y = x \leftrightarrow y = \underline{0}) \wedge \forall x \forall y (y \wedge x = x \leftrightarrow y = \underline{0})$
- A8 $\forall x_1 \forall x_2 \forall y \forall z (\text{sub}(x_1, y, z) \wedge \text{sub}(x_2, y, z)) = \text{sub}(x_1 \wedge x_2, y, z)$
- A9 $\neg \bar{a} = \bar{b}$, if a and b are distinct expressions.

A1-A4 describe the functions of concatenation, quotation and substitution by providing function values for specific entries. From these axioms one cannot derive (non-trivial) universally quantified principles and therefore axioms like the associative law for \wedge A5 are not derivable from A1-A4.

The concatenation of two expressions e_1 and e_2 is simply the expression e_1 followed by e_2 . For instance, $\neg\neg v$ is the concatenation of \neg and $\neg v$.

Therefore $\overline{\neg\neg v} = \overline{\neg} \wedge \overline{\neg v}$ is an instance of A2 as well as $\overline{\neg\neg v} = \overline{\neg} \wedge \overline{v}$.

Concatenating the empty string with any expression e gives again the same expression e . Therefore we have, for instance, $\overline{v} \wedge \underline{\quad} = \overline{v}$ as an instance of A2.

An instance of A3 is the sentence $q\overline{\overline{v}} = \overline{\overline{v}}$. Thus q describes the function that takes an expression and returns its quotation constant.

In A4 I have imposed the restriction that b must be a single symbol. This does not imply that the substitution function cannot be applied to complex expressions; just A4 does not say anything about the result of substituting a complex expression.

The reason for this restriction is that the result of substitution of a complex strings may be not unique. For instance, the result of substituting \neg for \wedge in $\wedge \wedge \wedge$ might be either $\wedge \neg$ or $\neg \wedge$. The problem can be fixed in several ways, but I do not need to substitute complex expressions in the following. Therefore I do not 'solve' the problem but avoid it by the restriction of b to a single symbol.

A1-A4 are already sufficient for proving the diagonalization Theorem 12.

A5 simplifies the reasoning with strings a great deal. Since $\mathcal{A} \vdash (x \wedge y) \wedge z = x \wedge (y \wedge z)$, that is, \wedge is associative by A5, I shall simply write $x \wedge y \wedge z$. for the sake of definiteness we can stipulate that $x \wedge y \wedge z$ is short for $(x \wedge y) \wedge z$ and similarly for more applications of \wedge .

I write $\mathcal{A} \vdash \varphi$ if and only if the formula φ is a logical consequence of the theory \mathcal{A} .

Example

$$\mathcal{A} \vdash \text{sub}(\overline{\overline{\neg}}, \overline{\neg}, \overline{\overline{\neg\neg}}) = \overline{\overline{\neg\neg\neg\neg\neg\neg}}$$

Example

$$\mathcal{A} \vdash \text{sub}(\overline{\overline{v = v \wedge \overline{v} = \overline{v}}}, \overline{v}, \overline{\overline{v_2}}) = \overline{\overline{v_2 = v_2 \wedge \overline{v} = \overline{v}}}$$

These axioms suffice for proving Gödel's celebrated diagonalization lemma.

Remark

Of course, there is no such cheap way to Gödel's theorems. Gödel showed that the functions sub and q (and further operations) can be defined in an arithmetical theory for numerical codes of expressions. To this end he proved that all recursive functions can be represented in a fixed arithmetical system. And then he proved that the operation of substitution etc. are recursive. This requires some work and ideas.

Diagonalization

The diagonalization function dia is defined in the following way:

Definition

$$\text{dia}(x) = \text{sub}(x, \bar{v}, q(x))$$

Remark

There are at least two ways to understand the syntactical status of dia . It may be considered an additional unary function of \mathcal{L} , and the above equation is then an additional axiom of \mathcal{A} .

Alternatively, one can conceive dia as a metalinguistic abbreviation, which does not form part of the language \mathcal{L} , but which is just short notation for a more complex expression. This situation will be encountered in the following frequently.

A lemma

Lemma

Assume $\varphi(v)$ is a formula not containing bound occurrences of v .
Then the following holds:

$$\mathcal{A} \vdash \overline{\text{dia}(\overline{\varphi(\text{dia}(v))})} = \overline{\varphi(\overline{\text{dia}(\overline{\varphi(\text{dia}(v))})})}$$

Proof.

In \mathcal{A} the following equations can be proved::

$$\begin{aligned} \overline{\text{dia}(\overline{\varphi(\text{dia}(v))})} &= \text{sub}(\overline{\varphi(\text{dia}(v))}, \bar{v}, \overline{q(\overline{\varphi(\text{dia}(v))})}) \\ &= \text{sub}(\overline{\varphi(\text{dia}(v))}, \bar{v}, \overline{\overline{\varphi(\text{dia}(v))}}) \\ &= \overline{\varphi(\overline{\text{dia}(\overline{\varphi(\text{dia}(v))})})} \end{aligned}$$

The diagonal lemma

Theorem (diagonalization)

If $\varphi(v)$ is a formula of \mathcal{L} with no bound occurrences of v , then one can find a formula γ such that the following holds:

$$\mathcal{A} \vdash \gamma \leftrightarrow \varphi(\bar{\gamma})$$

Proof.

Choose as γ the formula $\varphi(\text{dia}(\overline{\varphi(\text{dia}(v))}))$. Then one has by the previous Lemma:

$$\mathcal{A} \vdash \underbrace{\varphi(\text{dia}(\overline{\varphi(\text{dia}(v))}))}_{\gamma} \leftrightarrow \varphi(\underbrace{\overline{\varphi(\text{dia}(\overline{\varphi(\text{dia}(v))}))}}_{\gamma})$$

Formal Theories of Truth II
THE T-SENTENCES

Volker Halbach

17th February 2019

Inconsistency

I shall prove the inconsistency of some theories with the theory \mathcal{A} . 'inconsistent' always means 'inconsistent with \mathcal{A} '.

Since I did not fix the axioms of \mathcal{A} and admitted further axioms in \mathcal{A} , inconsistency results can be formulated in two ways. One can either say ' \mathcal{A} is inconsistent if it contains the sentence ψ ' or one says ' ψ is inconsistent with \mathcal{A} '.

The T-scheme

The first inconsistency result is the famous liar paradox. It is plausible to assume that a truth predicate T for the language \mathcal{L} satisfies the T-scheme

$$(3) \quad T\bar{\psi} \leftrightarrow \psi$$

for all sentences ψ of \mathcal{L} . This scheme corresponds to the scheme

'A' is true if and only if A,

where A is any English declarative sentence.

The liar in \mathcal{A}

Theorem (liar paradox)

The T-scheme $T\bar{\psi} \leftrightarrow \psi$ for all sentences ψ of \mathcal{L} is inconsistent.

Proof.

Apply the diagonalization theorem 12 to the formula $\neg Tv$. Then theorem 12 implies the existence of a sentence γ such that the following holds: $\mathcal{A} \vdash \gamma \leftrightarrow \neg T\bar{\gamma}$. Together with the instance $T\bar{\gamma} \leftrightarrow \gamma$ of the T-scheme this yields an inconsistency. γ is called the 'liar sentence'. ⊥

Tarski's theorem

Since the scheme is inconsistent such a truth predicate cannot be defined in \mathcal{A} , unless \mathcal{A} itself is inconsistent.

Corollary (Tarski's theorem on the undefinability of truth)

There is no formula $\tau(v)$ such that $\tau(\bar{\psi}) \leftrightarrow \psi$ can be derived in \mathcal{A} for all sentences ψ of \mathcal{L} , if \mathcal{A} is consistent.

Proof.

Apply the diagonalization theorem 12 to $\tau(v)$ as above. If $\tau(v)$ contains bound occurrences of v they can be renamed such that there are no bound occurrences of v . \neg

The scope of Tarski's theorem

It is not so much surprising that the axioms listed explicitly in Definition 4 do not allow for a definition of such truth predicate $\tau(v)$. According to Definition 4, however, \mathcal{A} may contain arbitrary additional axioms. Thus Tarski's Theorem says that adding axioms to \mathcal{A} that allow for a truth definition renders \mathcal{A} inconsistent.

Extending the language

Nevertheless one can add a new predicate symbol *which is not in \mathcal{L}* , and add $\text{True } \bar{\psi} \leftrightarrow \psi$ as an axiom scheme for all sentences of \mathcal{L} . In this case ψ cannot contain the symbol True and the diagonalization theorem 12 does not apply to $\text{True } v$ because it applies only to formulæ $\varphi(v)$ of \mathcal{L} .

Theorem

Assume that the language \mathcal{L} is expanded by a new predicate symbol True and all sentences $\text{True } \bar{\psi} \leftrightarrow \psi$ (for ψ a sentence of \mathcal{L}) are added to \mathcal{A} . The resulting theory is consistent if \mathcal{A} is consistent.

The theory of disquotation

Call the theory \mathcal{A} plus all these equivalences TB . Thus TB is given by the following set of axioms:

$$\mathcal{A} \cup \{\text{True } \bar{\psi} \leftrightarrow \psi : \psi \text{ a sentence of } \mathcal{L}\}$$

The proof

The idea for the proof is due to Tarski.

I shall show that a given proof of a contradiction \perp in the theory TB can be transformed into a proof of \perp in \mathcal{A} . In the given proof only finitely many axioms with True can occur; let

$$\text{True } \overline{\psi_0} \leftrightarrow \psi_0, \text{True } \overline{\psi_1} \leftrightarrow \psi_1, \dots, \text{True } \overline{\psi_n} \leftrightarrow \psi_n$$

be these axioms. $\tau(v)$ is the following formula of the language \mathcal{L} :

$$(v = \overline{\psi_0} \wedge \psi_0) \vee (v = \overline{\psi_1} \wedge \psi_1) \vee \dots \vee (v = \overline{\psi_n} \wedge \psi_n)$$

Obviously one has

$$\tau(\overline{\psi_0}) \leftrightarrow \psi_0$$

and similarly for all ψ_k ($k \leq n$).

Now replace everywhere in the given proof any formula True t , where t is any arbitrary term, by $\tau(t)$ and add above any former axiom True $\overline{\psi_k} \leftrightarrow \psi_k$ a proof of $\tau(\overline{\psi_k}) \leftrightarrow \psi_k$, respectively. The resulting structure is a proof in \mathcal{A} of the contradiction \perp .

Conservativity

The proof establishes a stronger result: Adding the T-sentences

$$\text{True } \bar{\psi} \leftrightarrow \psi$$

(ψ a sentence without True) to \mathcal{A} yields a conservative extension of \mathcal{A} :

Theorem

TB is conservative over \mathcal{A} . That is, If φ is a sentence without True that is provable in TB, then φ is already provable in \mathcal{A} only.

Proof.

Just replace \perp by φ in the proof above.

⊢

The proof shows that these T-sentences do not allow to prove any new 'substantial' insights. Works also with full induction.

Conservativity over logic

The T-sentences are not conservative over pure *logic*. The T-sentences prove that there are at least two different objects:

$$\text{True } \overline{\overline{\bar{V} = \bar{V}}} \leftrightarrow \bar{V} = \bar{V}$$

T-sentence

$$\bar{V} = \bar{V}$$

tautology

$$\text{True } \overline{\bar{V} = \bar{V}}$$

two preceding lines

$$\text{True } \overline{\neg \bar{V} = \bar{V}} \leftrightarrow \neg \bar{V} = \bar{V}$$

T-sentence

$$\neg \text{True } \overline{\neg \bar{V} = \bar{V}}$$

$$\overline{\bar{V} = \bar{V}} \neq \overline{\neg \bar{V} = \bar{V}}$$

Montague's paradox

Theorem (Montague's paradox [6])

The schema $T\bar{\psi} \rightarrow \psi$ is inconsistent with the rule $\frac{\psi}{T\bar{\psi}}$.

The rule $\frac{\psi}{T\bar{\psi}}$ is called NEC in the following.

Proof.

$\gamma \leftrightarrow \neg T\bar{\gamma}$	diagonalization
$T\bar{\gamma} \leftrightarrow \neg\gamma$	
$T\bar{\gamma} \rightarrow \gamma$	axiom
$\neg T\bar{\gamma}$	logic
γ	first line
$T\bar{\gamma}$	NEC

Another formulation of the T-sentences

Often the T-sentences are stated in the following way:

$$T\overline{\psi} \leftrightarrow \psi$$

where ψ must not contain T . It's thought that this is safe. But I don't trust that formulation anymore.

How not to state the T-sentences

Theorem

Assume \mathcal{L} contains a unary predicate symbol N (for necessity of some kind, let's say), and assume further:

- T $T\bar{\varphi} \leftrightarrow \varphi$ for all sentences φ of \mathcal{L} not containing T .
- N_1 $N\bar{\varphi} \rightarrow \varphi$ for all sentences φ of \mathcal{L} not containing N .
- N_2 Whenever $\mathcal{A} \vdash \varphi$, then also $\mathcal{A} \vdash N\bar{\varphi}$ for all sentences φ of \mathcal{L} not containing N .

Then \mathcal{A} is inconsistent.

$$\gamma \leftrightarrow \neg T \overline{N \overline{\gamma}}$$

diagonalisation

$$T \overline{N \overline{\gamma}} \leftrightarrow \neg \gamma$$

$$N \overline{\gamma} \rightarrow \neg \gamma$$

$$N \overline{\gamma} \rightarrow \gamma$$

(N1)

$$\neg N \overline{\gamma}$$

two previous lines

$$\neg T \overline{N \overline{\gamma}}$$

T

$$\gamma$$

first and last line

$$N \overline{\gamma}$$

(N2)

How not to state the T-sentences

Usually it is thought that typing is a remedy to the paradoxes. The example shows that this works only as long as typing is not applied to more than one predicate.

The result is the first of various paradoxes (*vulgo* inconsistencies) that arise from the interaction of predicates.

Summary

- Adding a new truth predicate to \mathcal{A} and axiomatising it by typed T-sentences yields a conservative extension of \mathcal{A} .
- The resulting theory TB does not prove generalisation such as

$\forall x(\text{Sent}(x) \rightarrow (\text{True } x \vee \text{True } \neg x))$ or

$\forall x\forall y(\text{Sent}(x) \wedge \text{Sent}(y) \rightarrow (\text{True } (x \wedge y) \leftrightarrow (\text{True } x \wedge \text{True } y)))$

- TB is not finitely axiomatisable.
- According to Tarski, a decent theory of truth should not only yield the T-sentences (and satisfy Convention T), but also prove those generalisations.
- 'Mixing' the T-sentences with axiomatisations of other notions such as necessity can lead to inconsistencies. So type restrictions don't solve all problems.

Liberalising the type restriction

There have been various proposals to lift the type restrictions on the T-sentences.

Motives:

- Eg the following T-sentence looks ok:

“Grass is red’ is not true’ is true iff ‘Grass is red’ is not true.

- A more liberal approach might help to regain deductive power.

However, one seems to be caught between Scylla and Charybdis: the typed truth predicate of TB is too weak, while the full unrestricted T-schema is too strong.

It seems reasonable to steer between the two extremes in the middle...

But there are other creatures as horrifying as deductive weakness and inconsistency, as McGee [5] has demonstrated.

Horwich's proposal

[...] we must conclude that permissible instantiations of the equivalence schema are restricted in some way so as to avoid paradoxical results. [...] Given our purposes it suffices for us to concede that certain instances of the equivalence schema are not to be included as axioms of the minimal theory, and to note that the principles governing our selection of excluded instances are, in order of priority: (a) that the minimal theory not engender 'liar-type' contradictions; (b) that the set of excluded instances be as small as possible; and—perhaps just as important as (b)—(c) that there be a constructive specification of the excluded instances that is as simple as possible.

Horwich 1990 p. 41f

Maximal consistent instances of schema T

So the aim is to find a set of sentences $T\bar{\varphi} \leftrightarrow \varphi$ such that

- The set is consistent.
- The set is maximal, ie no further sentences of the form $T\bar{\varphi} \leftrightarrow \varphi$ can be consistently added over \mathcal{A} .
- The set is recursively enumerable (?).

Maximal consistent instances of schema T

Theorem (McGee)

Let φ be some sentence, then there is a sentence γ such that

$$\mathcal{A} \vdash \varphi \leftrightarrow (T\bar{\gamma} \leftrightarrow \gamma)$$

Proof.

$$\mathcal{A} \vdash \gamma \leftrightarrow (T\bar{\gamma} \leftrightarrow \varphi)$$

diagonalisation

$$\mathcal{A} \vdash \varphi \leftrightarrow (T\bar{\gamma} \leftrightarrow \gamma)$$

propositional logic

Maximal consistent instances of schema T

McGee's observation spells disaster for Horwich's proposal.

Theorem (McGee)

- *If a consistent set of T-sentences is recursive, it's not maximal: by Gödel's first incompleteness theorem there will be an undecidable sentence φ , which is equivalent to a T-sentence.*
- *Maximal sets are too complicated. They can't be Π_1^0 or Σ_1^0 .*
- *There are many, in fact uncountably many different maximal consistent sets of T-sentences (if \mathcal{A} is consistent).*
- *Consistent sets of T-sentences can prove horrible results worse than any inconsistency.*

Strong instances of schema T

McGee's observation has its destructive uses, but it also has a neglected constructive side.

It's often assumed that an axiomatisation of truth by T-sentences is either inacceptably weak or inconsistent. McGee's theorem shows that this view is incorrect.

Assume you have a favourite axiomatisation of truth (say the KF axioms or the like). Let χ the conjunction of these axioms. Then McGee's theorem implies the existence of a T-sentence such that

$$\mathcal{A} \vdash \chi \leftrightarrow (\text{True } \bar{\gamma} \leftrightarrow \gamma)$$

Thus Davidson's theory, KF and so on can be finitely axiomatised by a single T-sentence.

The problem remains to tell a story why one should accept $T\bar{\gamma} \leftrightarrow \gamma$. If one justifies the acceptance of that T-sentence by appeal to your favourite theory, we have given up disquotationalism.

Strong instances of schema T

Gut feeling

- *Tarski's way of blocking the paradoxes is less damaging to the 'inductive' definition of truth than to the T-sentences as axioms.*
- *The T-sentences are as good as any axiomatic theory of truth, if the paradoxes are blocked in an appropriate way.*
- *We need to come up with a better method for sorting the good instances from the bad instances of schema T.*

To me it's still unclear whether it might be possible to defend a theory based on T-sentence which is not deductively weak.

Missing: maximal conservative sets of instances of schema T
Cieśliński [2]. 'Uniform' T-sentences and positive T-sentences. I don't have the tools available for treating them now. But there are well motivated and strong theories based on T-sentences.

Formal Theories of Truth IV
MORE BEASTS AND DRAGONS

Volker Halbach

17th February 2019

How things can go wrong

Paradox is not the same as mere inconsistency: there are many ways things can go wrong:

- The theory is inconsistent.
- The theory cannot be combined with another plausible theory. If a theory of future cannot be combined with the analogous theory of past truth, something is wrong.
- The theory is internally inconsistent: the theory proves that everything is true.
- The theory proves a false claim in the base language (ie in the language without the truth predicate).
- The theory has trivial models, eg, truth can be interpreted by the empty set.
- The theory is ω -inconsistent.

Generally, consistency proofs are good, but a full proof-theoretic analysis is better. Only such an analysis can prove that the theory doesn't contain any hidden paradoxes.

The constructive applications

On the next couple of slides I sketch some classical applications of diagonalisation.

Many of them can be turned into 'paradoxes'.

Gödel's first theorem

Ok, it isn't Gödel's incompleteness theorem, but it's very similar in structure:

Theorem (Gödel's first theorem)

Assume $\mathcal{A} \vdash \varphi$ if and only if $\mathcal{A} \vdash T\bar{\varphi}$ holds for all sentences. Then there is a sentence γ , such that neither γ itself nor its negation is derivable in \mathcal{A} except that \mathcal{A} itself is already inconsistent.

- | | | |
|------|--|-----------------|
| (4) | $\mathcal{A} \vdash \gamma \leftrightarrow \neg T\bar{\gamma}$ | diagonalisation |
| (5) | $\mathcal{A} \vdash \gamma$ | assumption |
| (6) | $\mathcal{A} \vdash T\bar{\gamma}$ | NEC |
| (7) | $\mathcal{A} \vdash \neg T\bar{\gamma}$ | (4) |
| (8) | $\mathcal{A} \vdash \neg \gamma$ | assumption |
| (9) | $\mathcal{A} \vdash T\bar{\gamma}$ | (4) |
| (10) | $\mathcal{A} \vdash \gamma$ | CONEC |

The liar again

Theorem

Assume $\mathcal{A} \vdash \varphi$ if and only if $\mathcal{A} \vdash T\bar{\varphi}$ holds for all sentences. Then the liar sentence is undecidable in \mathcal{A} , if \mathcal{A} is consistent.

Thus if the T-schema is weakened to a rule, the liar sentence must be undecidable. Thus theories (such as KF) containing NEC and deciding the liar sentence, cannot have CONEC.

The real incompleteness theorem

Gödel showed that a provability predicate $\text{Bew}(v)$ can be defined in a certain system of arithmetic corresponding to our theory \mathcal{A} . more precisely, he defined a formula $\text{Bew}(v)$

$$\mathcal{A} \vdash \psi \text{ if and only if } \mathcal{A} \vdash \text{Bew}(\overline{\psi})$$

holds for all formulæ ψ of \mathcal{L} if \mathcal{A} is ω -consistent. ω -consistency is a stronger condition than pure consistency.

A look at the second incompleteness theorem

The 'modal' reasoning leading to the second incompleteness theorem can be paraphrased in \mathcal{A} .

The second incompleteness theorem and Löb's theorem have been used to derive further paradoxes. I believe that most paradoxes involving self-reference can be reduced to Löb's theorem.

In particular, the incompleteness theorems yield more information on weakenings of the T-scheme and ways to block Montague's paradox.

Weaker reflection

Theorem

The scheme $\overline{\overline{TT\bar{\varphi}}} \rightarrow \varphi$ is inconsistent with NEC. The same holds for $\overline{\overline{TT\bar{\bar{\varphi}}}} \rightarrow \varphi$ etc.

Proof.

$$\mathcal{A} \vdash \gamma \leftrightarrow \neg \overline{\overline{TT\bar{\gamma}}}$$

$$\mathcal{A} \vdash \overline{\overline{TT\bar{\gamma}}} \rightarrow \gamma$$

assumption

$$\mathcal{A} \vdash \neg \overline{\overline{TT\bar{\gamma}}}$$

two preceding lines

$$\mathcal{A} \vdash \gamma$$

first line

$$\mathcal{A} \vdash T\bar{\gamma}$$

NEC

$$\mathcal{A} \vdash \overline{\overline{TT\bar{\gamma}}}$$

4

Internal inconsistency

Plain inconsistency is not the only way a system can fail to be acceptable. “Internal” inconsistency is almost as startling. Let \perp be some fixed logical contradiction, e.g., $\neg \neq \neg$. A theory is said to be internally inconsistent (with respect to T) if and only if $\mathcal{A} \vdash T\bar{\perp}$.

Theorem (Thomason [8])

The schema $\overline{TT\bar{\varphi}} \rightarrow \varphi$ is internally inconsistent with NEC.

Proof.

One runs the proof of Montague’s theorem in the scope of T . \dashv

The Löb derivability conditions

Let K be the following scheme:

$$(K) \quad T\overline{\varphi} \rightarrow \overline{\varphi} \rightarrow (T\overline{\varphi} \rightarrow T\overline{\varphi})$$

4 is the following scheme:

$$(4) \quad T\overline{\varphi} \rightarrow T\overline{T\overline{\varphi}}$$

K_4 contains NEC, K , 4 and all axioms of \mathcal{A} . K_4 has been thought to be adequate for necessity and, in some cases, for truth.

Remark

One can show that Gödel's provability predicate satisfies K_4 . NEC, K , 4 formulated for the provability predicate are known as Löb's derivability conditions. See [1] for more information.

Löb's theorem

Now I want to generalise the question: for which sentences can we have $T\bar{\varphi} \rightarrow \varphi$?

Theorem (Löb's theorem)

$$K_4 \vdash \overline{TT\bar{\varphi} \rightarrow \varphi} \rightarrow T\bar{\varphi}$$

The corresponding rule follows as well:

Theorem

If $K_4 \vdash T\bar{\varphi} \rightarrow \varphi$, then $K_4 \vdash \varphi$

Thus in the context of K_4 adding $T\bar{\varphi} \rightarrow \varphi$ makes φ itself provable.

Löb's theorem: the proof

Proof.

$\gamma \leftrightarrow (T\bar{\gamma} \rightarrow \varphi)$	diagonalization
$T\bar{\gamma} \rightarrow \overline{TT\bar{\gamma} \rightarrow \varphi}$	K
$T\bar{\gamma} \rightarrow \overline{TT\bar{\gamma}} \rightarrow T\bar{\varphi}$	K and NEC
$T\bar{\gamma} \rightarrow T\bar{\varphi}$	4
$(T\bar{\varphi} \rightarrow \varphi) \rightarrow (T\bar{\gamma} \rightarrow \varphi)$	
$(T\bar{\varphi} \rightarrow \varphi) \rightarrow \gamma$	first line
$\overline{T(T\bar{\varphi} \rightarrow \varphi) \rightarrow \gamma}$	NEC
$T\overline{(T\bar{\varphi} \rightarrow \varphi)} \rightarrow T\bar{\gamma}$	K
$T\overline{(T\bar{\varphi} \rightarrow \varphi)} \rightarrow T\bar{\varphi}$	line 4

Gödel's second theorem

Now fix a contradiction, for instance $\underline{0} \neq \underline{0}$ and call it \perp .

Theorem (Gödel's second theorem)

K_4 is inconsistent with $\neg T\perp$. Thus $K_4 \not\vdash \neg T\perp$ if K_4 is consistent.

There is also a formalized version of Gödel's second incompleteness theorem, which can easily be derived from Löb's theorem.

Theorem (Gödel's second theorem formalized)

$K_4 \vdash T\perp \vee \overline{\neg T\neg T\perp}$.

The dark side again

Now here is another paradox. I didn't know where else it should be put.

Very much like the paradox on how not to formalise the T-sentences is arises from the interaction of two predicates, viz two truth predicates: future and past truth.

Horsten and Leitgeb call it the 'no future' paradox.

No future: the language

Assume \mathcal{L} contains four predicates G , H , F and P . The intended reading of Gx is “ x always will be the case”, while Hx should be read as “ x always has been the case”. Similarly Fx is to be read as “ x will be the case at some point (in the future)”; finally Px stands for “ x has been the case at some point (in the past)”. G and H can easily be defined from F and P , respectively (or also vice versa). The four predicates correspond to the well known operators from temporal logic, the difference being that G and H are here predicates rather than operators.

No future: the axioms

The system K_t^* is given by the following axiom schemes for all sentences φ and ψ of the language \mathcal{L} . This means in particular that φ and ψ may contain the predicates G and H.

$$G1 \quad G\overline{\varphi} \rightarrow \overline{\psi} \rightarrow (G\overline{\varphi} \rightarrow G\overline{\psi})$$

$$H1 \quad H\overline{\varphi} \rightarrow \overline{\psi} \rightarrow (H\overline{\varphi} \rightarrow H\overline{\psi})^1$$

$$G2 \quad \varphi \rightarrow H\overline{F\overline{\varphi}}$$

$$H2 \quad \varphi \rightarrow G\overline{P\overline{\varphi}}$$

$$G3 \quad G\overline{\varphi} \leftrightarrow \neg F\overline{\neg\varphi}$$

$$H3 \quad H\overline{\varphi} \leftrightarrow \neg P\overline{\neg\varphi}$$

$$N \quad \frac{\varphi}{G\overline{\varphi}} \text{ and } \frac{\varphi}{H\overline{\varphi}} \text{ for all sentences } \varphi.$$

¹In [3] there is a typo in the formulation of this axiom: the last occurrence of H is a G in the original paper.

No future: the inconsistency

These axioms are analogues of axioms from temporal logic.

K_t^* is consistent (see [3]), but “internally” inconsistent, i.e., K_t^* proves that there is no future and no past.

Theorem (no future paradox, [3])

$$K_t^* \vdash H\perp \wedge G\perp.$$

Thus K_t^* claims that at all moments in the future \perp will hold. Since \perp is a contradiction, there cannot be any moment in the future. Therefore there is no future. Analogously, but less dramatically, there also has never been a moment in the past.

No future: the proof

I shall only prove that there is no future, i.e., $K_t^* \vdash G\perp$.

- (11) $K_t^* \vdash \gamma \leftrightarrow \overline{GP\neg\gamma}$ diagonalisation
 $K_t^* \vdash \neg\gamma \leftrightarrow \neg\overline{GP\neg\gamma}$
 $K_t^* \vdash \neg\gamma \rightarrow \gamma$ H2
- (12) $K_t^* \vdash \gamma$
- (13) $K_t^* \vdash \overline{GP\neg\gamma}$ from (??) and previous line
 $K_t^* \vdash H\bar{\gamma}$ N and (12)
 $K_t^* \vdash \neg P\neg\bar{\gamma}$ H3
- (14) $K_t^* \vdash \overline{G\neg P\neg\bar{\gamma}}$ N
- (15) $K_t^* \vdash G\perp$ (13), (14) and G1

The last line follows because we have $\overline{G\varphi \rightarrow (\neg\varphi \rightarrow \perp)}$ for all φ and, in particular, for $P\neg\bar{\gamma}$, by N.

No future: an inconsistency

In this framework one can assert that there is a future by saying that if φ will always be the case then φ will be the case at some time:

No future: an inconsistency

(FUT)

$$G\bar{\varphi} \rightarrow F\bar{\varphi}$$







Corollary ([3])

H₂, G₃, H₃, N and FUT together are inconsistent.

Proof.

One proves (13) and (14) as in the preceding Theorem and applies FUT to the latter in order to obtain $F\overline{\neg P\neg\gamma}$, which implies in turn $\neg G\overline{P\neg\gamma}$ by G₃ and is therefore inconsistent with (13). ⊥

Actually Horsten and Leitgeb [3] have proved the dual of this corollary.

-  George Boolos.
The Logic of Provability.
Cambridge University Press, Cambridge, 1993.
-  Cezary Cieśliński.
Deflationism, Conservativeness and Maximality.
Journal of Philosophical Logic, 36:695–705, 2007.
-  Leon Horsten and Hannes Leitgeb.
No Future.
Journal of Philosophical Logic, 30:259–265, 2001.
-  Paul Horwich.
Truth.
Basil Blackwell, Oxford, first edition, 1990.
-  Vann McGee.
Maximal consistent sets of instances of Tarski's schema (T).
Journal of Philosophical Logic, 21:235–241, 1992.
-  Richard Montague.

Syntactical Treatments of Modality, with Corollaries on Reflexion Principles and Finite Axiomatizability.

Acta Philosophica Fennica, 16:153–67, 1963.

Reprinted in [7, 286–302].



Richard Montague.

Formal Philosophy: Selected Papers of Richard Montague.

Yale University Press, New Haven and London, 1974.

Edited and with an introduction by Richmond H. Thomason.



Richmond H. Thomason.

A Note on Syntactical Treatments of Modality.

Synthese, 44:391–396, 1980.