NEW SESSION 2
Diagonalization and Axioms for Truth

Volker Halbach

Fudan

17th May 2012

Here is an informal version of diagonalisation due to Quine.

The quotation of an expression is the expression surrounded by quotation marks. So the quotation of

> *gjnw4 fwr*

is

> *'gjnw4 fwr'*

The quotation of

> *is not true*

is

> *'is not true'*

Here is an informal version of diagonalisation due to Quine.
The quotation of an expression is the expression surrounded by quotation marks. So the quotation of

gjnw4 fwr

is

'gjnw4 fwr'

The quotation of

is not true

is

'is not true'

Now consider the sentence:

liar

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now let’s look at what the following expression denotes:

‘followed by its own quotation is not true’ followed by its own quotation

It’s

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now consider the sentence:

liar

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now let’s look at what the following expression denotes:

‘followed by its own quotation is not true’ followed by its own quotation

It’s

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now consider the sentence:

liar

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now let’s look at what the following expression denotes:

‘followed by its own quotation is not true’ followed by its own quotation

It’s

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now consider the sentence:

liar

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Now let’s look at what the following expression denotes:

‘followed by its own quotation is not true’ followed by its own quotation

It’s

‘followed by its own quotation is not true’ followed by its own quotation is not true.

Hence the liar sentence claims about itself that it's not true and we have:

'followed by its own quotation is not true' followed by its own quotation is not true if and only if ' 'followed by its own quotation is not true' followed by its own quotation is not true' is not true

Hence we have a sentence L such that

L is and only if L is not true.

L is the liar sentence of course.

The trick works also with expressions other than 'is not true'.

The diagonal lemma from last time is only a generalisation of this trick.

Theorem (diagonalization)

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

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

The trick works also with expressions other than 'is not true'.

The diagonal lemma from last time is only a generalisation of this trick.

Theorem (diagonalization)

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

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

The trick works also with expressions other than ‘is not true’.

The diagonal lemma from last time is only a generalisation of this trick.

Theorem (diagonalization)

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

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

Why is this method of getting self-reference better than the method via labels?

- The method uses only basic syntactic operations.
- It does not depend on any empirical facts (important for paradoxes involving necessity)

Why is this method of getting self-reference better than the method via labels?

- The method uses only basic syntactic operations.
- It does not depend on any empirical facts (important for paradoxes involving necessity)

The T-scheme

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

$$N\overline{\psi} \leftrightarrow \psi \tag{1}$$

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 $N\overline{\psi} \leftrightarrow \psi$ *for all sentences* ψ *of* $\mathcal{L}$ *is inconsistent.*

Proof.

Apply the diagonalization theorem 1 to the formula $\neg N\mathrm{v}$. Then theorem 1 implies the existence of a sentence γ such that the following holds: $\mathcal{A} \vdash \gamma \leftrightarrow \neg N\overline{\gamma}$. Together with the instance $N\overline{\gamma} \leftrightarrow \gamma$ of the T-scheme this yields an inconsistency. γ is called the 'liar sentence'. $\dashv$

The liar in $\mathcal{A}$

Theorem (liar paradox)

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

Proof.

Apply the diagonalization theorem 1 to the formula $\neg N\mathrm{v}$. Then theorem 1 implies the existence of a sentence γ such that the following holds: $\mathcal{A} \vdash \gamma \leftrightarrow \neg N\overline{\gamma}$. Together with the instance $N\overline{\gamma} \leftrightarrow \gamma$ of the T-scheme this yields an inconsistency. γ is called the 'liar sentence'. $\dashv$

The use of formulae is not necessary. The method above produces an English sentence L such that

L if and only if L is not true

If we also have

L is true if and only if L is true.

we get

L is true if and only if L is not true

This is a contradiction (L can be neither true nor not true).

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 $\tau(\mathrm{v})$ such that $\tau(\overline{\psi}) \leftrightarrow \psi$ can be derived in $\mathcal{A}$ for all sentences ψ of $\mathcal{L}$, if $\mathcal{A}$ is consistent.

Proof.

Apply the diagonalization theorem 1 to $\tau(\mathrm{v})$ as above. If $\tau(\mathrm{v})$ contains bound occurrences of v they can be renamed such that there are no bound occurrences of v. $\dashv$

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 $\tau(\mathrm{v})$ such that $\tau(\overline{\psi}) \leftrightarrow \psi$ can be derived in $\mathcal{A}$ for all sentences ψ of $\mathcal{L}$, if $\mathcal{A}$ is consistent.

Proof.

Apply the diagonalization theorem 1 to $\tau(\mathrm{v})$ as above. If $\tau(\mathrm{v})$ contains bound occurrences of v they can be renamed such that there are no bound occurrences of v. $\dashv$

The scope of Tarski's theorem

It is not so much surprising that the axioms listed explicitly in Definition of $\mathcal{A}$ do not allow for a definition of such truth predicate $\tau(\mathrm{v})$. 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.

Consequences of Tarski's theorem

- Given a modest amount of syntax theory (or the like), truth cannot be defined in a consistent theory.
- The usual definitional theories (correspondence, coherence, pragmatic) are affected by this theorem.
- The truth is never simple.

Consequences of Tarski's theorem

- Given a modest amount of syntax theory (or the like), truth cannot be defined in a consistent theory.
- The usual definitional theories (correspondence, coherence, pragmatic) are affected by this theorem.
- The truth is never simple.

Consequences of Tarski's theorem

- Given a modest amount of syntax theory (or the like), truth cannot be defined in a consistent theory.
- The usual definitional theories (correspondence, coherence, pragmatic) are affected by this theorem.
- The truth is never simple.

The Theorem applies only if we are trying to define a truth predicate satisfying

'A' is true if and only if A.

for all sentence A of the language.

We might well be able to define a truth predicate that satisfies the equivalences for many but not all A.

But in philosophy we often would like to make very global claims without restricting our scope to a specific vocabulary.

The Theorem applies only if we are trying to define a truth predicate satisfying

'A' is true if and only if A.

for all sentence A of the language.

We might well be able to define a truth predicate that satisfies the equivalences for many but not all A.

But in philosophy we often would like to make very global claims without restricting our scope to a specific vocabulary.

The Theorem applies only if we are trying to define a truth predicate satisfying

'A' is true if and only if A.

for all sentence A of the language.

We might well be able to define a truth predicate that satisfies the equivalences for many but not all A.

But in philosophy we often would like to make very global claims without restricting our scope to a specific vocabulary.

Mathematical logicians have more or less given up on the notion of global truth and are happy with relativized notions of truth.

However, if such a predicate cannot be define, we can add one just by adding the equivalences as axioms.

That is we find a new predicate symbol T and use all sentence $T\overline{\psi} \leftrightarrow \psi$ as axioms, where ψ is a sentence from the original language (without the symbol T).

So, informally speaking, we have

'A' is true if and only if A.

as axioms for all sentences with the expression 'is true'.

However, if such a predicate cannot be define, we can add one just by adding the equivalences as axioms.

That is we find a new predicate symbol T and use all sentence $T\overline{\psi} \leftrightarrow \psi$ as axioms, where ψ is a sentence from the original language (without the symbol T).

So, informally speaking, we have

'*A*' is true if and only if *A*.

as axioms for all sentences with the expression 'is true'.

However, if such a predicate cannot be define, we can add one just by adding the equivalences as axioms.

That is we find a new predicate symbol T and use all sentence $T\overline{\psi} \leftrightarrow \psi$ as axioms, where ψ is a sentence from the original language (without the symbol T).

So, informally speaking, we have

‘A’ is true if and only if A.

as axioms for all sentences with the expression ‘is true’.

The theory of disquotation

The theory TB (‘Tarski biconditionals’) is given by all axioms of our syntax theory $\mathcal{A}$ and all axioms

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

for sentence ψ of the original language without T.

Think of TB as your theory of syntax (which does not contain the epxression ‘is true’); and assume you add ‘is true’ to this language fragment together with the axioms

‘A’ is true if and only if *A*.

for all sentence *A* of the original language.

The theory of disquotation

The theory TB ('Tarski biconditionals') is given by all axioms of our syntax theory $\mathcal{A}$ and all axioms

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

for sentence ψ of the original language without T.

Think of TB as your theory of syntax (which does not contain the epxression 'is true'); and assume you add 'is true' to this language fragment together with the axioms

'A' is true if and only if A.

for all sentence A of the original language.

The theory *TB* is ‘disquotational’ or ‘deflationist’. This and related theories play an important role in Quine’s, Davidson’s, and Horwich’s theories of truth.

It’s disquotational because the truth predicate ‘cancels out’ quotation marks.

According to some philosophers, the only purpose of the truth predicate is to cancel out quotation marks (or similar devices); they think there is nothing more to say about truth than just *TB*.

The theory *TB* is 'disquotational' or 'deflationist'. This and related theories play an important role in Quine's, Davidson's, and Horwich's theories of truth.

It's disquotational because the truth predicate 'cancels out' quotation marks.

According to some philosophers, the only purpose of the truth predicate is to cancel out quotation marks (or similar devices); they think there is nothing more to say about truth than just *TB*.

The theory *TB* is ‘disquotational’ or ‘deflationist’. This and related theories play an important role in Quine’s, Davidson’s, and Horwich’s theories of truth.

It’s disquotational because the truth predicate ‘cancels out’ quotation marks.

According to some philosophers, the only purpose of the truth predicate is to cancel out quotation marks (or similar devices); they think there is nothing more to say about truth than just *TB*.

Results:

- TB is consistent (so adding these disquotational axioms doesn't cause any problems; but we need to be careful to avoid 'interaction' paradoxes).
- TB is conservative over a basic theory of syntax such as $\mathcal{A}$. That means that no new sentences without the truth predicate follow from the axioms for truth; thus this theory of truth doesn't give us any new insights that are not truth-theoretic.

These results do not imply that a truth predicate given by the TB axioms is useless.

The truth predicate of TB can still be used to express generalizations. For instance, form the assumption

Everything the pope says is true.

one can derive the conclusion

If the pope says 'Frogs taste good', then frogs taste good.

The truth axioms of TB are used for this argument.

These results do not imply that a truth predicate given by the TB axioms is useless.

The truth predicate of TB can still be used to express generalizations. For instance, form the assumption

> *Everything the pope says is true.*

one can derive the conclusion

> *If the pope says 'Frogs taste good', then frogs taste good.*

The truth axioms of TB are used for this argument.

Conservativity over logic

The Tarski biconditionals are not conservative over pure *logic*. The T-sentences prove that there are at least two different objects, because one can prove that there is an object that is true and another object that is false.

The discussion on conservativeness was started by Horsten (1995), Shapiro (1998), Ketland (1999) with replies by Field (1999).

Conservativity over logic

The Tarski biconditionals are not conservative over pure *logic*. The T-sentences prove that there are at least two different objects, because one can prove that there is an object that is true and another object that is false.

The discussion on conservativeness was started by Horsten (1995), Shapiro (1998), Ketland (1999) with replies by Field (1999).

Many philosophers have complained that the theory TB is too weak. It doesn't give us all the consequences we would like to derive from a good theory of truth.

I give an example.

Many philosophers have complained that the theory TB is too weak. It doesn't give us all the consequences we would like to derive from a good theory of truth.

I give an example.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

. . .

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

...

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

...

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

...

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

...

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

. . .

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Using just Tarski biconditionals we can show (I'm now finally abbreviating 'if and only if' as 'iff'):

'Snow is white' is not true iff 'snow is not white' is true.

'Frogs taste good' is not true iff 'frogs don't taste good' is true.

'All mammals are cows' is not true iff 'not all mammals are cows' is true.

...

You see the pattern:

'A' is not true iff 'not-A' is true.
where *A* is some sentence without the truth predicate.

Hence we would like to prove the generalisation:

For all sentences: the sentence is not true iff the negation of the sentence is true.

But we cannot prove this from the Tarski-biconditionals: in any given argument we can use only finitely many of them, but the generalisation requires all of them (Tarski gave a formal proof and rejected TB because of its deductive weakness).

More notation

In order to prove the result that TB doesn't prove those generalisations, I need more axioms.

Sent(x) is a unary predicate, $\dot{\neg}$ a unary function symbol. I assume that Sent(x) represents the property of being a sentence of $\mathcal{L}$, $\dot{\neg}$ represents the function that takes a sentence and returns its negation:

Additional Axiom

$\mathcal{A} \vdash Sent(\overline{\varphi})$ *iff* φ *is a sentence of* $\mathcal{L}$.

Additional Axiom

$\mathcal{A} \vdash \dot{\neg}\overline{\varphi} = \overline{\neg\varphi}$

More notation

In order to prove the result that TB doesn't prove those generalisations, I need more axioms.

Sent(x) is a unary predicate, $\underset{\cdot}{\neg}$ a unary function symbol. I assume that Sent(x) represents the property of being a sentence of $\mathcal{L}$, $\underset{\cdot}{\neg}$ represents the function that takes a sentence and returns its negation:

Additional Axiom

$\mathcal{A} \vdash \mathit{Sent}(\overline{\varphi})$ *iff* φ *is a sentence of* $\mathcal{L}$.

Additional Axiom

$\mathcal{A} \vdash \underset{\cdot}{\neg}\overline{\varphi} = \overline{\neg\varphi}$

No generalisations

I can now formulate and proof Tarski's complaint:

Theorem

$TB \nvdash \forall x(Sent(x) \to (Tx \vee T\dot{\neg}x))$ *(assuming that $\mathcal{A}$ is consistent).*

Proof

Assume otherwise. Then there is a proof of $\forall x(\mathrm{Sent}(x) \rightarrow (Tx \vee T\dot{\neg}x))$ in from a finite subtheory S of TB. Only finitely many T-sentences can be in S. Let

$$T\overline{\psi_0} \leftrightarrow \psi_0, T\overline{\psi_1} \leftrightarrow \psi_1, \ldots, T\overline{\psi_n} \leftrightarrow \psi_n$$

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

$$\big((\mathrm{v} = \overline{\psi_0} \wedge \psi_0) \vee (\mathrm{v} = \overline{\psi_1} \wedge \psi_1) \vee \ldots (\mathrm{v} = \overline{\psi_n} \wedge \psi_n)\big) \wedge (\mathrm{v} = \overline{\psi_1} \vee \ldots \vee \mathrm{v} = \overline{\psi_n})$$

As above, $T\mathrm{v}$ can be interpreted as $\tau(\mathrm{v})$.

If χ is none of the $\psi_0, \ldots, \psi_n$, we have $\mathcal{A} \vdash \neg\tau(\overline{\chi}) \wedge \neg\tau(\dot{\neg}\overline{\chi})$.

Proof

Assume otherwise. Then there is a proof of $\forall x(\mathrm{Sent}(x) \to (Tx \vee T\underset{\cdot}{\neg}x))$ in from a finite subtheory S of TB. Only finitely many T-sentences can be in S. Let

$$T\overline{\psi_0} \leftrightarrow \psi_0, T\overline{\psi_1} \leftrightarrow \psi_1, \ldots, T\overline{\psi_n} \leftrightarrow \psi_n$$

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

$$\big((\mathrm{v} = \overline{\psi_0} \wedge \psi_0) \vee (\mathrm{v} = \overline{\psi_1} \wedge \psi_1) \vee \ldots (\mathrm{v} = \overline{\psi_n} \wedge \psi_n)\big) \wedge (\mathrm{v} = \overline{\psi_1} \vee \ldots \vee \mathrm{v} = \overline{\psi_n})$$

As above, $T\mathrm{v}$ can be interpreted as $\tau(\mathrm{v})$.

If χ is none of the $\psi_0, \ldots, \psi_n$, we have $\mathcal{A} \vdash \neg\tau(\overline{\chi}) \wedge \neg\tau(\underset{\cdot}{\neg}\overline{\chi})$.

Davidson made another objection: the theory TB cannot be finitely axiomatized. So how can we ever learn the truth predicate?

I don't agree with Davidson about this, but TB is in fact not finitely axiomatizable.

Davidson made another objection: the theory TB cannot be finitely axiomatized. So how can we ever learn the truth predicate?

I don't agree with Davidson about this, but TB is in fact not finitely axiomatizable.

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

$$\forall x(\mathrm{Sent}(x) \rightarrow (Tx \vee T\dot{\neg}x)) \text{ or}$$
$$\forall x \forall y\big(\mathrm{Sent}(x) \wedge \mathrm{Sent}(y) \rightarrow (T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty))\big)$$

- 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.

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

$$\forall x(\mathrm{Sent}(x) \rightarrow (Tx \vee T\dot{\neg}x)) \text{ or}$$

$$\forall x \forall y\big(\mathrm{Sent}(x) \wedge \mathrm{Sent}(y) \rightarrow (T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty))\big)$$

- 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.

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

$$\forall x(\mathrm{Sent}(x) \rightarrow (Tx \vee T\dot{\neg}x)) \text{ or}$$

$$\forall x \forall y\big(\mathrm{Sent}(x) \wedge \mathrm{Sent}(y) \rightarrow (T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty))\big)$$

- 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.

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

$$\forall x(\mathrm{Sent}(x) \rightarrow (Tx \vee T\dot{\neg}x)) \text{ or}$$

$$\forall x \forall y\big(\mathrm{Sent}(x) \wedge \mathrm{Sent}(y) \rightarrow (T(x \dot{\wedge} y) \leftrightarrow (Tx \wedge Ty))\big)$$

- 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.

Philosophical moral

The truth predicate of TB may have its merits: it allows one to axiomatise certain generalisations finitely Horwich (1998) Halbach (1999). But it doesn't prove the generalisations Tarski expected from a decent theory of truth.

Moreover, TB has been criticised, because the object-/metalanguage distinction seems to restrictive.

Philosophical moral

The truth predicate of TB may have its merits: it allows one to axiomatise certain generalisations finitely Horwich (1998) Halbach (1999). But it doesn't prove the generalisations Tarski expected from a decent theory of truth.

Moreover, TB has been criticised, because the object-/metalanguage distinction seems to restrictive.

Liberalising the type restriction

There have been various proposals to lift the type restrictions on the T-sentences, ie. to admit also sentences.

'A' is true iff A

where *A* may contain the truth predicate.

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.

Liberalising the type restriction

There have been various proposals to lift the type restrictions on the T-sentences, ie. to admit also sentences.

'A' is true iff A

where *A* may contain the truth predicate.

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 (1992) has demonstrated.

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 (1992) has demonstrated.

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 (1992) 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

McGee (1992) proved that this proposal leads to problems: it doesn't single out a single set of Tarski biconditionals, and, even worse, these theories can have disastrous consequences.

*References

George Boolos. *The Logic of Provability*. Cambridge University Press, Cambridge, 1993.

Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96:533–540, 1999.

Volker Halbach. Disquotationalism and infinite conjunctions. *Mind*, 108:1–22, 1999.

Leon Horsten. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In P. Cortois, editor, *The Many Problems of Realism*, volume 3 of *Studies in the General Philosophy of Science*, pages 173–187. Tilburg University Press, Tilburg, 1995.

Paul Horwich. *Truth*. Basil Blackwell, Oxford, first edition, 1990.

Paul Horwich. *Truth*. Oxford University Press, Oxford, second edition edition, 1998. first edition 1990.

Jeffrey Ketland. Deflationism and Tarski's paradise. *Mind*, 108:69–94, 1999.

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