What a revenge-free solution to the liar paradox can and can't look like

Andrew Bacon

February 2012

The T-Schema:

> 'ϕ' is true in English if and only if ϕ.

The problem:

> For some sentence: $L =$ 'L is not true'
>
> By the T-schema: 'L is not true' is true if and only if L is not true
>
> By Leibniz' law: L is true if and only if L is not true
>
> By classical logic we may conclude any sentence from this whatsoever.

The meaning schema:

> 'ϕ' means that ϕ in English.

Assuming that every sentence means at most one thing (in fact, you need only assume that everything a sentence means is materially equivalent) the meaning schema is inconsistent. (You can define a truth predicate as $\exists p(\ulcorner\phi\urcorner$ means that p and p).)

1 Revenge

Suppose we have some classification of sentences into 'healthy' and 'diseased'. Assuming the disease in question is supposed to be whatever causes the T-schema to fail, we may formulate two natural constraints on a theory of healthiness and truth:

1. The theory should assert that all healthy sentences satisfy the T-schema.
2. The theory itself shouldn't be diseased.

It is therefore natural to want a theory of the 'disease' to contain the following principle:

SRT. $H(\ulcorner\phi\urcorner) \rightarrow (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$

> (If 'ϕ' is healthy, then 'ϕ' is true if and only if ϕ.)

Let γ be the sentence 'γ is either unhealthy or untrue.' Unfortunately, assuming only SRT, we can prove both γ and that 'γ is unhealthy'. In other words, any theory containing SRT. can prove its own theorems to be unhealthy. Furthermore, any theory containing SRT and a necessitation principle is inconsistent:

N. If $\vdash \phi$ then $\vdash H(\ulcorner\phi\urcorner)$

> (If you can prove ϕ, you can prove that 'ϕ' is healthy.)
>
> Consider a sentence $\ulcorner\gamma\urcorner$, which says: either $\ulcorner\gamma\urcorner$ is unhealthy or untrue.
>
> Suppose, for contradiction, that $\ulcorner\gamma\urcorner$ is healthy. Then by the restricted T-schema, SRT, $\ulcorner\gamma\urcorner$ would be true if and only if it were either untrue or unhealthy ($Tr(\ulcorner\gamma\urcorner) \leftrightarrow (\neg H(\ulcorner\gamma\urcorner) \vee \neg Tr(\ulcorner\gamma\urcorner))$. But this can only happen if $\ulcorner\gamma\urcorner$ is unhealthy! So $\ulcorner\gamma\urcorner$ is not healthy after all, contradiction.
>
> Therefore, $\ulcorner\gamma\urcorner$ is unhealthy. Thus, $\ulcorner\gamma\urcorner$ is either unhealthy or untrue. But this is just what $\ulcorner\gamma\urcorner$ says. So we have a proof of γ. By necessitation, N., we have that $\ulcorner\gamma\urcorner$ is healthy. Contradiction.

Two things to note. Firstly, you can make this argument rigorous in an arithmetical setting. Secondly, the above argument can be reformulated so that the only rule of inference used is modus ponens (as it stands the argument uses 'proof by contradiction' – a rule which is sometimes rejected by classical logicians.)

The argument above also establishes that for any theory containing SRT (and enough arithmetic) there is some sentence, γ, such (i) γ is part of the theory and (ii) $\neg H(\ulcorner\gamma\urcorner)$, the claim that $\ulcorner\gamma\urcorner$ is unhealthy, is part of the theory.

2 A non-linguistic theory

Following the literature on vagueness, it is natural to formulate a non-linguistic theory of healthiness with an operator, Δp, pronounced 'it's determinate that p'.[1] From this one can define another operator $\nabla p := \neg\Delta p \wedge \neg\Delta\neg p$ pronounced 'it's indeterminate whether p.'

The claim that it's determinate that snow is white is no more a claim about language than the claim that it's necessary that snow is white. Concerning tense operators, Prior writes:

> When a sentence is formed out of another sentence or other sentences by means of an adverb or conjunction, it is not about those other sentences, but about whatever they are themselves about.

Without the meaning schema ('ϕ' means that ϕ) there is no general method of paraphrasing an expression of the form [operator]ϕ with a predicate attributing something to the sentence 'ϕ'.[2] Having a primitive operator, allowing us to talk directly about whether ϕ, affords us an expressive advantage.

The following minimal theory of determinacy and truth is consistent.

K $\Delta(\phi \rightarrow \psi) \rightarrow (\Delta\phi \rightarrow \Delta\psi)$

[1] Alternatively one could use a determinacy predicate, D, applying to propositions, and a 'that p' subnective for denoting propositions. 'It's determinate that p' can then be formalised as D(that p). This difference is mostly immaterial but it does require the consistency arguments to be modified.

[2] For more mundane reasons: let τ be a German sentence expressing the proposition that τ is true in German. This could be achieved by letting τ be 'τ ist wahr' for example. The fact that it's indeterminate whether τ is true in German seems to be a fact about the German language, having nothing to do with English. Any claim about the English sentence 'τ is true in German' is at best indirectly about the German language.

T $\Delta\phi \to \phi$

$\Delta\mathsf{Nec}$ If $\vdash \phi$ then $\vdash \Delta\phi$

RT $(\Delta\phi \vee \Delta\neg\phi) \to (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$

How does the revenge paradox manifest itself in this system? We may consider a sentence, δ, which says of itself that it is not determinately true: $\delta \leftrightarrow \neg\Delta Tr(\ulcorner\delta\urcorner)$. From this and the above principles we may show:

$$\neg\Delta\Delta\delta$$

Revenge results in higher order indeterminacy instead of inconsistency.

If an operator O satisfies the above axioms, so does the operator OO, OOO, etc. This raises two natural questions:

1. What is the intended interpretation of Δ?
2. Can we say something about the conceptual role of Δ to narrow its possible interpretations down?

I address question 1. in other work, in which Δ is explicitly defined in terms of the notion of coherent rational preferences. However, we may still give something like an implicit definition of Δ by outlining its role in a theory of rational propositional attitudes. For example, that one shouldn't care intrinsically about the indeterminate: if the strongest determinate proposition p entails is identical to the strongest determinate proposition q entails, you should be indifferent between p and q whenever p and q are maximally specific propositions. Another possible constraint, suggested by Hartry Field, is that indeterminacy requires your credences to be non-additive; for example, if you're certain that it's indeterminate whether p then your credences in p and in $\neg p$ should be 0. In what follows I want to concentrate on the attitudes of acceptance and rejection, and the permissibility of certain speech acts.

Assuming you have all the relevant information possible available to you:

> Accept p iff it's determinate that p.
>
> Reject p iff it's not determinate that p.

Pragmatics. Assuming you have as much relevant information available to you as possible:

> It's permissible to assert that p iff it's determinate that p.

Consequence: assuming that the above biconditional is determinate (I asserted it!) then:

> If it's indeterminate whether it's determinate that p, it's indeterminate whether it's permissible to assert that p.

(The fact that I stated these principles on assertability, acceptance and rejection using *biconditionals* instead of just conditionals allows us show that the iterations of Δ don't determinately satisfy the right conceptual role, even though they obey the four axioms.)

2.1 Can a theory describe its own pragmatics?

Certain non-classical theories of truth have difficulties containing their own theory of assertability and rational acceptance and rejection (see, e.g., Priest 1987, Soames 1999, Field 2008, Beall 2011) as do some classical theories (Feferman 1991, Maudlin 2004.) Field, for example, has a theory of determinacy with similar properties to the above one. But for Field, indeterminacy at any order licenses rejection, and you may only assert things which are determinate at all orders. For example, Field thinks that it's determinate that you should reject the revenge liar and all of its strengthenings ('I'm not determinately^n true' for iterations n.)

For Field, the assertability operator is not coextensive with any iteration of the determinacy operator (or indeed any definable operator) so his theory does not contain it's own account of assertability.

Secondly, for Field each iteration of the determinacy operator does a better job of characterising the paradoxes, but none of them do so perfectly. The current theory evades both these features.

In summary

- The current theory can contain its own account of acceptance and assertability.
- It makes a single distinction between the healthy (determinate) and unhealthy (indeterminate) propositions. Like many distinctions, this distinction has indeterminate instances (such as δ.)

3 Technical stuff (not discussed in talk)

The four principles I listed serve as a very basic theory for reasoning about determinacy and truth. I have shown that they are consistent, and furthermore consistent with a richer theory of determinacy and truth. Call the following theory DFS.

PA. Peano arithmetic including full induction (i.e. the induction schema may take formulae containing the truth predicate and the determinacy operator.)

ΔNec If $\vdash \phi$ then $\vdash \Delta\phi$

K $\Delta(\phi \to \psi) \to (\Delta\phi \to \Delta\psi)$

T $\Delta\phi \to \phi$

BF $\forall x \Delta\phi \to \Delta\forall x\phi$

RT $(\Delta\phi \vee \Delta\neg\phi) \to (Tr(\ulcorner\phi\urcorner) \leftrightarrow \phi)$

At. $\forall x(At(x) \to (Tr(x) \leftrightarrow Ver(x)))$

$\to$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x\dot{\to}y) \leftrightarrow (Tr(x) \to Tr(y)))$

$\vee$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x\dot{\vee}y) \leftrightarrow Tr(x) \vee Tr(y)))$

$\wedge$. $\forall x\forall y(Sent(x) \wedge Sent(y) \to (Tr(x\dot{\wedge}y) \leftrightarrow Tr(x) \wedge Tr(y)))$

$\forall$. $\forall x(Sent(x(\bar{0}/v)) \to (Tr(\dot{\forall}vx) \leftrightarrow \forall y Tr(x[\dot{y}/v])))$

$\neg$. $\forall x(Sent(x) \to (Tr(\dot{\neg}x) \leftrightarrow \neg Tr(x)))$

Δ. $\forall x(Sent(x) \to (Tr(\dot{\Delta}x) \leftrightarrow \Delta Tr(x)))$

Nec If $\vdash \phi$ then $\vdash Tr(\ulcorner\phi\urcorner)$

Conec If $\vdash Tr(\ulcorner\phi\urcorner)$ then $\vdash \phi$

The system DFS is consistent, and, in fact, doesn't prove any false arithmetical sentences. If you restrict induction to its arithmetical instances it proves no new arithmetical theorems. It's also consistently augmentable with propositional quantification (and full second order logic.)[3]

Theorem 3.1. *The following principles are inconsistent if added to DFS*

B $\phi \to \Delta\neg\Delta\neg\phi$

4 $\Delta\phi \to \Delta\Delta\phi$

5 $\neg\Delta\phi \to \Delta\neg\Delta\phi$

Theorem 3.2. *BF + T + ΔNec + RT + Δ., are ω-inconsistent.*

Theorem 3.3. *The following theories have standard models*

1. DFS - BF - Conec - $\forall$.
2. DFS - Δ. - Conec - $\forall$.
3. DFS - Δ. - BF - $\forall$.

[3] DFS is closely related to the truth theory FS (which it contains as a subsystem), which is, by a result due to McGee, ω-inconsistent. Theorem 3.3 identifies three ω-consistent subsystems of DFS.