A Formal Classification of Pathological Satisfaction Classes

Alexander Jones

University of Bristol

Bristol-München Conference on Truth and Rationality, 11th June 2016

A Formal Classification of Pathological Sentences

I will be looking at satisfaction classes over Peano Arithmetic.

I want to discuss what makes some satisfaction classes 'pathological' and what makes others acceptable.

In particular, I want to give a formal criterion of this in terms of Robinson's notion of semantic entailment for nonstandard sentences.

Table of Contents

Introduction

The typed theory of truth CT^- - also known as $\mathrm{PA}(S)^-$ - has a semantic interpretation in the form of satisfaction classes - sets of Gödel codes of sentences satisfying the compositional clauses.
Every satisfaction class S is adequate in the sense that for standard sentence φ:

$$M \vDash \varphi \text{ if and only if } (M, S) \vDash S(\ulcorner \varphi \urcorner, c)$$

Some satisfaction classes contain nonstandard sentences which are intuitively false, however, such as:

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee \ldots \vee 0 = 1)))$$

Motivation

One of the reasons that the theory of truth CT^- is viewed as an unattractive theory is because the satisfaction classes it produces are not satisfactory.

> *Should we rest happy with $\mathrm{PA}(S)^-$ then? That would be a rather hasty conclusion ... the following generalisation is not provable in $\mathrm{PA}(S)^-$: take a false sentence α, produce a disjunction of an arbitrary length with α as the only disjunct, and the result of your operation will also be false [Cieśliński, 2010, Page 329].*

There is an active research programme in removing pathologies from the theory of satisfaction classes, but a good question to ask is what exactly *are* the pathological sentences. Which sentences are pathological and what is the reason for this?

Satisfaction Class

Definition of a Satisfaction Class

A set $S \subseteq M \times M$ is a satisfaction class if $(\varphi, c) \in S$ if and only if $M \vDash Form(\varphi)$ and c is the code of an assignment of free variables to elements of M. Further, $(M, S) \vDash S(\varphi, c)$ if and only if one of the following conditions holds:

1. $CT1(\varphi, c)$: $\exists m, n[Term(m) \wedge Term(n) \wedge \varphi = (n = m) \wedge Val(n, c) = Val(m, c)]$
2. $CT2(\varphi, c)$: $\exists \alpha, \beta[Form(\alpha) \wedge Form(\beta) \wedge \varphi = (\alpha \wedge \beta) \wedge (S(\alpha, c) \wedge S(\beta, c))]$
3. $CT3(\varphi, c)$: $\exists \alpha, \beta[Form(\alpha) \wedge Form(\beta) \wedge \varphi = (\alpha \vee \beta) \wedge (S(\alpha, c) \vee S(\beta, c))]$
4. $CT4(\varphi, c)$: $\exists \psi[Form(\psi) \wedge \varphi = \neg\psi \wedge \neg S(\psi, c)]$
5. $CT5(\varphi, c)$: $\exists \psi[Form(\psi) \wedge \varphi = \exists y \psi \wedge \exists b S(\psi, c[\frac{y}{b}])]$
6. $CT6(\varphi, c)$: $\exists \psi[Form(\psi) \wedge \varphi = \forall y \psi \wedge \forall b S(\psi, c[\frac{y}{b}])]$

Key Theorems

Lachlan's Theorem

If $M \vDash \mathrm{PA}$ and M has a satisfaction class S, then M is recursively saturated [Kotlarski, 1991, Theorem 3].

KKL's Theorem

If $M \vDash \mathrm{PA}$ and M is countable and recursively saturated, then M has a satisfaction class S [Kotlarski, 1991, Theorem 2].

Theorem

If $M \vDash \mathrm{PA}$ and M has a satisfaction class S, then M has $2^{\aleph_0}$-many such satisfaction classes [Kotlarski, 1991, Theorem 1].

Theorem

If $M \vDash \mathrm{PA}$ and S is a satisfaction class for M which is closed under Δ_0-induction, then $(M, S) \vDash S(\mathit{Con}(\mathrm{PA}))$ [Cieśliński, 2010, Page 332].

Pathologies in the Literature

Examples of pathological sentences that can be found within the literature are:

1. The sentence:

$$\delta_a^{(0\neq 0)}, \text{ for nonstandard } a \text{ and}$$

$$\delta_0^{(0\neq 0)} \text{ is } (0 \neq 0) \text{ and}$$

$$\delta_{n+1}^{(0\neq 0)} \text{ is } (\delta_n^{(0\neq 0)} \vee \delta_n^{(0\neq 0)}) \text{ for all } n \in M \text{ [Cieśliński, 2010, Page 327].}$$

2. For a nonstandard number a:

$$\exists x_0, x_1, ..., x_a[0 \neq 0] \text{ [Engström, 2002, Page 56].}$$

3. For a nonstandard number a:

$$\exists x_0 \forall x_1 \exists x_2 ... \forall x_{2a-1} \exists x_{2a}[\varphi] \leftrightarrow \neg\varphi \text{ [Engström, 2002, Page 56].}$$

Further Examples of Pathological Sentences

1. For a nonstandard a, the sentence:

$$(0 = 1 \vee (0 = \underline{2} \vee (0 = \underline{3} \vee (\ldots \vee (0 = \underline{a-1} \vee 0 = \underline{a}) \ldots))))$$

2. For a nonstandard a, the sentence:

$$\mho^0 = 0_{2a}, \text{ where}$$

$$\mho^0 = 0_0 \text{ is } 0 = 0, \text{ and}$$

$$\mho^0 = 0_{n+1} \text{ is } \neg\mho^0 = 0_n \text{ for all } n \in M.$$

3. There are also 'true' sentences which can be false in a satisfaction class. Let φ be a true sentence and a be a nonstandard number:

$$*_a^\varphi, \text{ where}$$

$$*_0^\varphi \text{ is } (\varphi \wedge \varphi), \text{ and}$$

$$*_{n+1}^\varphi \text{ is } (*_n^\varphi \wedge *_n^\varphi) \text{ for each } n \in M.$$

Informal Identification

The examples given can all be true in a satisfaction class, but are false according to our intuition.

The sentences where this behaviour is considered especially problematic is when the truth-value of these sentences seems so obvious to us from our perspective, and should never be viewed as true.

As a rough, heuristic, definition, it is exactly these sentences which are the pathologies.

Notable Common Features

It is important to note that in all of the examples given we believe that we can understand them, despite their nonstandard nature - in contrast to an arbitrary nonstandard sentence.

Further, none of the sentences depend on a particular model considered. They look false in every model of PA that is considered.

Lastly, they all contain a nonstandard number of connectives, but the individual clauses are (mostly) standard.

Naive Criteria

Since all the examples contain a nonstandard number of connectives, one might conclude that a pathological sentence is one with a nonstandard number of connectives.

- This is too strong, however. Consider the sentence:

$$(0 = 1 \wedge (0 = 1 \wedge (0 = 1 \wedge \ldots \wedge 0 = 1)))$$

The sentences are equivalent (in some sense) to ones which only contain a standard number of connective. One might conclude that those sentences which are pathological are sentences which are equivalent (in some sense) to a standard-finite sentence.

- This criterion is too weak, however. Consider the example:

$$(0 = 1 \vee (0 = \underline{2} \vee (0 = \underline{3} \vee \ldots \vee (0 = \underline{a-1} \vee 0 = \underline{a}) \ldots))))$$

Cieśliński's First Criterion

Cieśliński considers two different notions of pathological sentences. The first is:

> *"A class P of pathological cases will consist of all the sentences φ (in the sense of the model M) such that for some natural number n, $M \vDash Tr_n(\neg\varphi)$, with '$Tr_n(.)$' being an appropriate partial truth predicate." [Cieśliński, 2010, Page 331]*

We cannot class all of the pathologies in this way though, since some pathologies have complexity beyond Σ_n for any $n \in \mathbb{N}$. For example:

$$\exists x_0 \forall x_1 \exists x_2 ... \forall x_{2a-1} \exists x_{2a} [\varphi] \leftrightarrow \neg\varphi$$

Another worry is that this will include some sentences which, although of complexity Σ_n, are so complicated that we cannot understand them at all.

Cieśliński's Second Criterion

The second notion of pathological sentence that Cieśliński considers is that:

> *"The set of pathologies would be simply the set of all sentences disprovable in first order logic." [Cieśliński, 2010, Page 331]*

This says that if a formalised provability in propositional logic predicate is introduced, and this predicate states that a sentence σ is disprovable, but there is a satisfaction class which believes σ is true, then the sentence σ is pathological.
This criterion, again, does not include all pathological sentences, however. We can consider the sentence:

$$\neg\exists x_0, x_1, ..., x_a \bigwedge_{0\leqslant i,j\leqslant a} x_i \neq x_j$$

which is false in a nonstandard model M, but also not disprovable in PA.

Truth-Contingent Sentences

One thing which is to be noticed about pathological sentences is that all of them may be either true or false in a specific satisfaction class.
I shall use the phrase *truth-contingent* to define these sentences, those that can be either true or false in a satisfaction class.

Definition of Truth-Contingent

A sentence φ in ${}^*\mathscr{L}_A(M)$ is truth-contingent if and only if there are satisfaction classes S_1 and S_2 such that $(M, S_1) \vDash S_1(\ulcorner\varphi\urcorner, c)$ and $(M, S_2) \vDash \neg S_2(\ulcorner\varphi\urcorner, c)$.

Whilst all pathological sentences are truth-contingent, the converse should not hold. There are sentences which are truth-contingent, but their truth values are not at all obvious (or the sentences even readable by us) that are not pathological.

Informal Proposal

The informal proposal for my classification is that the common element in pathological sentences is that they violate an intuitive equivalence schema.

We think that for a given pathological sentence φ, it is the case that $M \models^\star \neg\varphi$ for some notion of semantic entailment.

It is this guiding principle of truth that we appeal to when unhappy with the pathological sentences.

Is one able to come up with a formal notion of $\models^\star$ which allows this view to be fleshed out? Yes, Robinson has introduced one for nonstandard sentences, which I will now define.

Simple Terms and Formulas

Definition of a Simple Term and Formula

A term or formula φ of ${}^*\mathscr{L}_A(M)$ is defined to be simple by induction on the complexity of φ.

- A term φ is simple if it is the result of a finite number of applications of functions with a finite number of arguments.
- An atomic formula φ is simple if it consists of a relation in ${}^*\mathscr{L}_A(M)$ with a finite number of arguments which are all filled by simple terms.
- A simple atomic formula φ is a simple wff.
- The negation of a simple wff is simple.
- The repeated conjunction or disjunction of a (potentially infinite) set of simple wffs is simple.
- A wff obtained from a (potentially infinite) number of quantifiers over a simple wff is simple. [Robinson, 1963, Page 104]

Rank of a Wff

Definition of the Rank of a Wff

The rank of a simple wff φ of ${}^*\mathscr{L}_A(M)$, rank(φ), is defined by induction on the complexity of φ.

- If φ is an atomic wff, then rank(φ) $= 0$.
- If φ is the negation of a wff ψ, then rank(φ) $=$ rank(ψ) $+ 1$.
- If φ is the repeated conjunction of wffs $\psi_1, \psi_2, ..., \psi_a$ (assuming WLOG none of these are conjuncts themselves), then rank(φ) $=$ max$\{$rank$(\psi_i) + 1 : i = 1, 2, ..., a\}$.
- If φ is the repeated disjunction of wffs $\psi_1, \psi_2, ..., \psi_a$ (assuming WLOG none of these are disjuncts themselves), then rank(φ) $=$ max$\{$rank$(\psi_i) + 1 : i = 1, 2, ..., a\}$.
- If φ is the quantification of a wff ψ (assuming WLOG this is not immediately bound by quantifiers itself), then rank(φ) $=$rank(ψ) $+ 1$. [Robinson, 1963, Pages 104-5]

Definition of $\models^\star$

Definition of $M \models^\star \varphi$

$M \models^* \varphi$ is defined for a simple wff φ of ${}^*\mathscr{L}_A(M)$ which has finite rank by induction on the rank of φ.

- If $\text{rank}(\varphi) = 0$, then φ can be evaluated in the standard way and we write $M \models^* \varphi$ if and only if $M \models \varphi$.
- If $\text{rank}(\varphi) = n$ and φ is of the form $\neg\psi$, then we write $M \models^* \varphi$ if and only if $M \not\models^* \psi$.
- If φ has rank n and is a conjunction of simple wffs $\psi_1, \psi_2, ..., \psi_a$ of rank $< n$ then $M \models^* \varphi$ if and only if $M \models^* \psi_i$ for all i from 1 to a.
- If φ has rank n and is a disjunction of simple wffs $\psi_1, \psi_2, ..., \psi_a$ of rank $< n$ then $M \models^* \varphi$ if and only if $M \models^* \psi_i$ for some i from 1 to a. [Robinson, 1963, Pages 106-7]

Definition of $\models^\star$

Definition of $\models^\star$ Continued

- Suppose φ has rank n and is of the form $q\psi$ where ψ is simple wff of rank $< n$ with no immediate quantifiers and q is a (potentially infinite) string of quantifiers. We take the Skolemised form of $q\psi$ and denote this by $\xi(x_1, x_2, ..., x_p, f_1(\bar{y_1}), f_2(\bar{y_2}), ..., f_q(\bar{y_r}))$, where $f_i(\bar{y_j})$ are each of the Skolem functions. The formula ξ has rank $< n$. We say that $M \models^* \varphi$ if and only if $M \models^* \xi(x_1, x_2, ..., x_p, f_1(\bar{y_1}), f_2(\bar{y_2}), ..., f_q(\bar{y_r}))$ for all possible substitutions.[Robinson, 1963, Pages 106-7]

Formal Proposal of Pathological Sentence

With this notion of semantic entailment for nonstandard sentences, $\models^\star$ in place we are now able to define what it means to be a pathological sentence.

Definition of Pathological Sentence

A sentence φ of ${}^*\mathscr{L}_A(M)$ is pathological if it is a simple wff of finite rank such that there is a satisfaction class S for M such that $(M, S) \models S(\ulcorner\varphi\urcorner, c)$, but $M \models^* \neg\varphi$. We shall call a satisfaction class which exhibits such behaviour a pathological satisfaction class.

Intuitive Attraction

This definition is certainly intuitively attractive. If a sentence violates one of the most basic principles of truth, then it is certainly pathological. At no point do we want to accept a nonstandard sentence as internally true, if we can tell that it is externally false.

The condition that all sentences considered are simple wffs of finite rank is similarly desirable. This provides a formal categorisation of our desideratum that sentences we cannot understand from our perspective are not treated as pathological.

This definition allows us to consider sentences containing an infinite number of connectives, but not those which have such a complicated structure that it appears we cannot even describe them.

Covers Problematic Examples

This definition also ensures that the examples considered are treated pathological. For example, consider the sentence:

$$\exists x_0 \forall x_1 \exists x_2 \ldots \forall x_{2a-1} \exists x_{2a} [\dagger_b]$$

where $\dagger_b$ is the sentence:

$$(0 = 1 \vee (0 = 1 \vee (0 = 1 \vee (\ldots \vee (0 = 1 \vee 0 = 1) \ldots))))$$

where there are $b \in M \backslash \mathbb{N}$ disjuncts $0 = 1$.

This is because the sentence is a simple wff of finite rank (since $0 = 1$ is a simple wff of finite rank) and the definition of $\vDash^\star$ ensures that this sentence is evaluated as false according to a nonstandard model M.

Further Attractive Features

There are attractive features of the definition of $\models^\star$ which means that the good facets of Cieśliński's definitions are retained, whilst avoiding their problems.

The notion of $\models^\star$ adheres to the truth of a partial truth predicate Tr_n for all simple wffs of finite rank which are Σ_n, as $\models^\star$ adheres to the Tarski conditions for truth. It also evaluates and assigns good values to sentences of complexity beyond Σ_n, however.

Further, Robinson has shown that that all tautologies of first order logic are true in $\models^\star$ [Robinson, 1963, Pages 106-7] which means that logical truth is retained. The definition also encompasses truths of the model and not just first order logic, however.

Good definition

Therefore, I propose that this is a good definition of the notion of pathological sentence.

It captures the intuition that a pathological sentence is false and violating some kind of equivalence schema, by utilising a notion of external falsity for nonstandard sentences.

It does not apply to sentences which are so complicated in their nonstandardness that we cannot even comprehend them.

Lastly, it evaluates the typical examples of pathological sentences as false and thus adheres to good test cases.

Interesting Questions

Is this a sufficient definition of pathology?

Let f be a function symbol and consider the sentence

$$f(f(...f(0))) = f(f(...f(0)))$$

where there are a nonstandard number of iterations. Is this a pathology which should also be eliminated?

Can we improve upon the notion of $\models^\star$?

Is there a good way to improve upon this notion of semantic entailment $\models^\star$ so that it covers more nonstandard sentences?

How conservative are satisfaction classes closed under $\models^\star$?

What are the model/proof-theoretic consequences of closing a satisfaction class for a model under this notion of $\models^\star$? Perhaps one arrives at $\Delta_0\text{–}\mathrm{PA}(S)$?

References I


Cieśliński, C. (2010).
Deflationary truth and pathologies.
J. Philos. Logic, 39(3):325–337.


Engström, F. (2002).
Satisfaction classes in nonstandard models of first-order arithmetic.
Department of Mathematics, Chalmers University of Technology and Göteborg University, no: 2002:24. Department of Mathematics, Chalmers University of Technology,.


Kotlarski, H. (1991).
Full satisfaction classes: a survey.
Notre Dame J. Formal Logic, 32(4):573–579.


Robinson, A. (1963).
On languages which are based on non-standard arithmetic.
Nagoya Math. J., 22:83–117.