Werk

Label: Article

Jahr: 1982

PURL: https://resolver.sub.uni-goettingen.de/purl?316342866_0023|log63

Kontakt/Contact

Digizeitschriften e.V.
SUB Göttingen
Platz der Göttinger Sieben 1
37073 Göttingen

info@digizeitschriften.de

COMMENTATIONES MATHEMATICAE UNIVERSITATIS CAROLINAE
23,4 (1982)

A LOGICAL ANALYSIS OF THE TRUTH-REACTION PARADOX

Kamila BENDOVÁ and Petr HÁJEK

Abstract: A logical paradox that originated in discussion on Artificial Intelligence (AI) is analyzed by means of arithmetization of metamathematics. The purpose is to isolate places where the apparent paradox disappears when notions occurring in it are understood as formal arithmetical notions and, secondly, to show how double use of self-reference in the paradox can be formalized.

Key words: Paradox, self-reference, Peano arithmetic, artificial intelligence.

Classification: Primary 03F30
Secondary 03B45, 68G99

---

§ 1. Introduction. The paradox in question was first formulated by Cherniavsky [1] in context of discussion on the difference between the reasoning of the man and the machine. His presentation is rather hard to understand; fortunately, the paradox was simplified and clarified by Havel [4] who called the paradox the truth-reaction paradox. Havel's aim was to contribute to the discussion mentioned above. In this paper we disregard completely the context of AI, take the paradox as it stands and look at it through the eyes of mathematical logic. This means that we replace intuitive notions involved in the paradox by formal notions (which can be done in various ways) and try either to show that a particular formal theory is inconsistant or to isolate

places where the would-be proof of a contradiction is not a proof and cannot be converted to a proof. This seems to be the professional duty of a mathematical logician when facing a paradox. (Needless to say, such an analysis does not mean that the paradox, formulated in intuitive terms, is meaningless or uninteresting.) We shall present such an analysis and show two possibilities of formalization inside formal arithmetic. The first one, based on the notions of provability, is simpler, but disregards the fact that the paradox uses two sorts of self-reference: first (usual) self-referential sentences (sentences referring to themselves) and, second, self-referential proofs (proofs referring to themselves). This second kind of self-reference seems to be new and the classical diagonalization lemma of Gödel-Feferman does not enable us to construct self-referential proofs. (The recursion theorem is the appropriate means for this.) The analysis results in a construction of a partial recursive function that formalizes attempts at constructing a proof of a contradiction; we show that all attempts fail.

The reader is assumed to know the elements of arithmetization of mathematics and of recursion theory. Knowledge of the papers by Enderton [2] and Smoryński [5] from the Handbook of Mathematical Logic is more than sufficient for the present discussion.

§ 2. The Paradox. In this part the Cherniavsky-Havel Truth-reaction paradox will be briefly described. (For details see [1],[4].)

Consider an informal theory with two binary predicates E and T, two axioms A1, A2 and an "evidence rule" Ev.

T(x,y) (usually written xTy) describes the experimental situation where some subject (either a human or a computer) accepts the text y as true after analyzing the text x. E(x,y) (written xEy) means that x includes evidence for y.

A1. $xEy \rightarrow xTy$
(If x includes evidence for y then the subject accepts y when knowing x.)

A2. $xE(xTy \rightarrow \neg y) \rightarrow \neg(xTy)$
(The subject does not want to be in contradiction with reality, more precisely: the subject will never accept a statement if this very act of accepting would make the statement false.)

Ev If you, after being exposed to x, feel convinced of y then xEy.
The particular case important for us is the following:
If x contains a proof of y then xEy. Havel does not specify the notion of proof involved here; he merely says "what is a proof for us (e.g. for the reader) should be an evidence for the subject, too."

The truth-reaction paradox follows.

1. Let d denote the text of this proof (lines 1. through 8.)
2. $s \equiv \neg(dTs)$ (self-referential definition of a sentence s)
3. $dTs \rightarrow \neg s$ (directly from 2.)
4. $dE(dTs \rightarrow \neg s)$ (1.-3. is a proof of 3. - thus 4. follows by Ev)
5. $\neg(dTs)$ A2 applied to 4.
6. s by 2. and 5.

7. dEs (1.-6. is a proof of s -
7. follows by Ev)

8. dEs & ¬(dTs) (5. and 7. put together)

And 8. contradicts A1.

In Havel s theory variables range over <u>texts</u> (they might be called text variables); formulas are particular texts and if x is a text and y is a formula then both xTy and xEy are formulas, thus particular texts. Note that proofs (from some axioms) are also particular texts. This suggests treating Havel's theory as a kind of propositional calculus or, better, text calculus with usual formation rules for propositional formulas enriched as follows:

(1) Any finite sequence of symbols is a text

(2) If x is a text and y is a formula then both xTy and xEy are formulas.

(T and E may be understood as some <u>modalities.</u>)

Consider a simplified particular case where T and E are identified and made independent of the first variable. Thus variables are propositional variables and T becomes necessity $\Box$

(A1) is trivialized,

(A2) obtains the form $\Box(\Box A \longrightarrow \neg A) \longrightarrow \neg \Box A$

Ev says "if A is provable then infer $\Box A$", which may be understood as necessitation: from A infer $\Box A$.

One further assumes

(A0) all propositional tautologies and
the existence of a particular propositional formula being a fixed-point of the modality $\neg\Box$:

(Fp) $q \equiv \neg \Box q$.

Theorem 1. Axioms (A0)-(A2),(Fp) together with modus ponens and necessitation form a contradictory modal theory.

(The paradoxical proof above amounts directly to a proof of a contradiction.)

Remark 1. This shows, in particular, that (A2) in the present form is not provable in the modal system G of [6].

§ 3. Arithmetical interpretations. Our general approach is as follows: We let variables of Havel's theory range over finite sequences of symbols of the language of Peano arithmetic PA, hereafter called PA-texts. The notion of provability involved in (Ev) will be made precise as meaning PA-provability (provability in PA). For each PA-text s we have its code (formalization) $\ulcorner s \urcorner$ which is a particular term describing s in PA.

An interpretation of Havel's theory will be any pair $\varepsilon(x,y)$, $\tau(x,y)$ of PA-formulas having (at most) two free variables; τ interprets T and ε interprets E. Havel's axioms become axiom schemes: For any PA-texts d, s, we have

(A1$^{\times}$) $\varepsilon(\ulcorner d \urcorner, \ulcorner s \urcorner) \longrightarrow \tau(\ulcorner d \urcorner, \ulcorner s \urcorner)$

(A2$^{\times}$) $\varepsilon(\ulcorner d \urcorner, \ulcorner \tau(\ulcorner d \urcorner, \ulcorner s \urcorner) \longrightarrow \neg s \urcorner) \longrightarrow \neg \tau(\ulcorner d \urcorner, \ulcorner s \urcorner)$

(Note that for each text s, $\neg s$ is also a text; if a is a formula then $\neg s$ is a formula.)

Let $\mathcal{T}$ be a theory containing PA. The interpretation is sound in $\mathcal{T}$ if (A1$^{\times}$) and (A2$^{\times}$) are provable in $\mathcal{T}$ (for each choice of d, s) and if the following holds (the rule Ev):

If d is a PA-proof of a formula s then $\mathcal{T} \vdash \varepsilon(\ulcorner d \urcorner, \ulcorner s \urcorner)$.

We shall exhibit some interpretations sound in some theories and not sound (ill) in some others. In each case we shall show how

the construction of the paradoxical would-be proof is formalizable and isolate the exact place where the exact place where the formalization is not a proof of contradiction.

To close this section, let us recall some facts about PA used in the sequel (see [4]):

Fact 1 (Diagonalization lemma). Let χ be a PA-formula with one free variable. Then there exists a sentence φ such that

$$PA \vdash \varphi \equiv \chi(\ulcorner \varphi \urcorner).$$

Fact 2 (about Σ_1-sentences). Every true closed Σ_1 PA-formula is a theorem of PA. (A PA-formula is Σ_1 if it has the form $(\exists x)\psi$ where all quantifiers in ψ are bounded, see [4].)

Fact 3. PA is consistent since it has a model (the standard model of natural numbers).

§ 4. An analysis based on provability. In the first interpretation we ignore the self-reference to d and identify T and E as one unary predicate. The predicate T (= E) must be interpreted in accordance with the evidence rule, i.e. if there is a PA-proof of a then $T(\ulcorner s \urcorner)$ must be provable. Therefore we interpret T as Pr(x) where Pr(x) is the standard predicate of provability in PA, i.e. for each PA-sentence φ we have

$$PA \vdash \varphi \quad \text{iff} \quad PA \vdash Pr(\ulcorner \varphi \urcorner).$$

More generally, if $\mathcal{T}$ is an axiomatized extension of PA then $Pr_{\mathcal{T}}(x)$ denotes the standard provability predicate for $\mathcal{T}$; thus Pr(x) is the same as $Pr_{PA}(x)$.

Thus this interpretation is sound in a theory $\mathcal{T}$ if we have the following:

(A2) $\mathcal{T} \vdash \mathrm{Pr}(\ulcorner \mathrm{Pr}(\ulcorner \varphi \urcorner) \to \neg\, \varphi \urcorner) \to \neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner)$

(Ev) If there is a proof of φ in PA then $\mathcal{T} \vdash \mathrm{Pr}(\ulcorner \varphi \urcorner)$.

The <u>paradox</u> translates to the following sequence of PA-formulas

2′. $\varphi \equiv \neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner)$ where φ is a particular sentence of PA such that $\mathrm{PA} \vdash \varphi \equiv \neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner)$.

3′. $\mathrm{Pr}(\ulcorner \varphi \urcorner) \to \neg\, \varphi$

4′. $\mathrm{Pr}(\ulcorner \mathrm{Pr}(\ulcorner \varphi \urcorner) \to \neg\, \varphi \urcorner)$

5′. $\neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner)$

6′. φ

7′. $\mathrm{Pr}(\ulcorner \varphi \urcorner)$

8′. $\neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner) \,\&\, \mathrm{Pr}(\ulcorner \varphi \urcorner)$

<u>Theorem 2</u>. The interpretation of T and E by Pr is not sound w.r.t. PA.

<u>Proof</u>. The axiom (A1) is provable in PA due to identification of T and E. (Ev) is true in PA by Fact 2, since Pr(x) is a Σ_1-formula.

Let us now suppose that (A2) is provable for every formula ψ. This means

$$\mathrm{PA} \vdash \mathrm{Pr}(\ulcorner \mathrm{Pr}(\ulcorner \psi \urcorner) \to \neg\, \psi \urcorner) \to \neg\, \mathrm{Pr}(\ulcorner \psi \urcorner)$$

for every ψ, especially for the sentence φ for which

$$\mathrm{PA} \vdash \varphi \equiv \neg\, \mathrm{Pr}(\ulcorner \varphi \urcorner).$$

Then in PA all steps in the Paradox are provable and we have a proof of contradiction on PA which is not possible (Fact 3.). Thus the axiom (A2) is not provable for every formula ψ.

<u>Theorem 3</u>. The interpretation of T and E by Pr is sound w.r.t. PA + $\mathrm{Con}_{\mathrm{PA}}$ where $\mathrm{Con}_{\mathrm{PA}}$ is the sentence formally expressing consistency of PA.

Proof. Note that if φ is a PA-sentence then

$$PA + Con_{PA} \vdash \neg Pr(\ulcorner\varphi\urcorner) \vee \neg Pr(\ulcorner\neg\varphi\urcorner).$$

We want to prove

$$PA + Con_{PA} \vdash Pr(\ulcorner Pr(\ulcorner\varphi\urcorner) \longrightarrow \neg\varphi\urcorner) \longrightarrow \neg Pr(\ulcorner\varphi\urcorner).$$

We have

$PA \vdash Pr(\ulcorner Pr(\ulcorner\varphi\urcorner) \longrightarrow \neg\varphi\urcorner) \longrightarrow (Pr(\ulcorner Pr(\ulcorner\varphi\urcorner)\urcorner) \longrightarrow Pr(\ulcorner\neg\varphi\urcorner))$

$PA \vdash Pr(\ulcorner\varphi\urcorner) \longrightarrow Pr(\ulcorner Pr(\ulcorner\varphi\urcorner)\urcorner)$

by the properties of Pr(x).

Thus we have

$PA, Con_{PA}, Pr(\ulcorner Pr(\ulcorner\varphi\urcorner) \longrightarrow \neg\varphi\urcorner), Pr(\ulcorner\varphi\urcorner) \vdash Pr(\ulcorner\neg\varphi\urcorner)$

which implies provability of $(A2^{\times})$.

Corollary (Gödel second incompleteness theorem). $PA \nvdash Con_{PA}$.

Immediate from Theorem 1 and Theorem 2.

Remark 2. Comparison with a proof of second Gödel's incompleteness theorem, e.g. in [5] makes obvious that the present proof is by no means a new proof of this celebrated result; it merely shows that the analyzed paradox - in the present interpretation - involves the same reasoning as that used in the proof of the second Gödel's Theorem.

Fact 4. Consider the theory PA + A2 where A2 is the axiom schema as above. We have $PA + A2 \vdash Con_{PA}$.

Proof. Let φ be a sentence such that

$$PA \vdash \varphi \equiv \neg Pr(\ulcorner\varphi\urcorner);$$

then $PA \vdash Pr(\ulcorner Pr(\ulcorner\varphi\urcorner) \longrightarrow \neg\varphi\urcorner)$ by Fact 2.

and thus

$$PA, A2 \vdash \neg Pr(\ulcorner\varphi\urcorner)$$

i.e.

$$PA, A2 \vdash (\exists x)(Fml(x) \& \neg Pr(x))$$

which is a sentence equivalent to Con_{PA}.

Remark 3. We saw in the proof of Theorem 1 that the sequence (2′)-(8′) is not a sequence of formulas provable in PA: we could not use (A2). Now, (A2) is provable in $PA+Con_{PA}$ (and $PA+Con_{PA}$ is consistent); filling the details makes (2′)-(6′) into a $(PA+Con_{PA})$-proof. But this does not entitle us to conclude $Pr_{PA}(\ulcorner\varphi\urcorner)$, but only that $Pr_{PA+Con}(\ulcorner\varphi\urcorner)$. Thus this is the place where the "proof of contradiction" collapses. (In fact we obtain a $(PA+Con_{PA})$-proof of $Pr_{PA+Con}(\ulcorner\varphi\urcorner) \& \neg Pr_{PA}(\ulcorner\varphi\urcorner)$.)

Remark 4. The reader can easily see that if we change our notion of soundness of (τ, ε) by postulating

(Ev′) If d is a $(PA+Con_{PA})$-proof of a then $\mathfrak{T} \vdash \varepsilon(\ulcorner d\urcorner, \ulcorner s\urcorner)$ then the interpretation of T and E by $Pr_{PA+Con_{PA}}(x)$ is not sound w.r.t. $PA+Con_{PA}$; similarly for other theories $\mathfrak{T} \supseteq PA$.

Similarly, the reader may verify that if Pr_{PA} is taken for ε (to interpret Havel's E) and $\tau(x)$ is any formula of the language of PA such that $PA \vdash (\forall x)(Pr_{PA}(x) \longrightarrow \tau(x))$ then the corresponding instance of A2 is unprovable in PA. (If we had such a proof then the paradox would yield a proof of contradiction in PA.)

§ 5. An analysis based on the notion of proof. The interpretation of Havel's E (and T) presented in the preceding section does not depend on the first argument of E (of T) and therefore the diagonalization over proofs is disregarded. Here we shall imitate the paradox more closely.

First realize that it is trivially impossible for a text

to contain itself as a proper subtext; the reference of a proof to itself is only possible by means of an appropriate description of a proof and not by quoting the whole proof in a part of itself.

We could identify descriptions of finite sequences with nullary Turing machines. If such machine d converges and produces an output $\{d\} = s$ then it is a successful description of s, if it diverges ($\{d\}$ is undefined) then d fails to describe a symbol. More generally, the description may have some finitely many steps, say 9. Such a description may be best understood as a unary Turing machine d for which we are interested only in values $\{d\}(0),\{d\}(1),\ldots,\{d\}(8)$. If all these values are defined then d successfully describes the word $s = \{d\}(0)* \ldots .* \{d\}(8)$ (where $*$ is the sign of concatenation); if d fails to describe a word in 9 steps we may at least ask whether d has succeeded to describe a word in (say) three steps, i.e. whether $\{d\}(0)$, $\{d\}(1)$, $\{d\}(2)$ all converge and if so, we can investigate the word $\{d\}(0)*\{d\}(1)*\{d\}(2)$.

We shall understand the paradoxical "proof" as a description of a binary Turing machine h that processes (mentions) each stepwise description d of a proof (successful or failing) and tries to describe a new proof (in 9 steps)

$$\{h\}(d,0),\{h\}(d,1),\ldots \{h\}(d,8).$$

By Recursion Theorem, we then investigate an arbitrary d such that

$$\{h\}(d,i) \cong \{d\}(i)$$

for each i (= 0,1,... 8). This will mean that the proof described by $\{d\}(0)$, $\{d\}(1)$,... mentions itself.

Havel's predicate E (identified with T throughout) is most

naturally interpreted by postulating that $\varepsilon(x,y)$ says x is a sequence and contains a proof of y, i.e. y is one of formulas (not necessarily the least one) occurring in x:

$Prv_{PA}(x,y) \equiv$ x is a sequence $\&(\exists w)$ (w is a subsequence of x $\&$ $\&$ w is a PA-proof of y).

Convention. d"n denotes $\{d\}(0) * \ldots * \{d\}(n-1)$; this value exists if and only if all values $\{d\}(0), \ldots \{d\}(n-1)$ exist. $Prv_{PA}(x"y,z)$ is assumed to mean: for some $y_1 \leqq y$, $x"y_1$ exists, say, $x"y_1 = w$, and $Prv_{PA}(w,z)$. Note that for each particular n, $Prv_{PA}(x"\bar{n},z)$ is a Σ_1-formula.
(Here $\bar{n}$ is the n-th numeral, i.e. the term $0\overbrace{'\ldots'}^{\text{n-times}}$.)

Construction. We define a Turing machine h by describing its behavior for arbitrary first argument d and for second argument i = 0,1,... 8. $\{h\}(d,i)$ for $i \geqq 9$ is irrelevant (say is 0). The definition parallels steps in the Truth-reaction paradox, the first step being postponed: it will consist in an application of the Recursion Theorem.

i = 0: Given d, the machine h constructs a self-referential sentence s such that

$$PA \vdash s \equiv \neg\, Prv_{PA}(\ulcorner d \urcorner"9, \ulcorner s \urcorner)$$

and outputs a PA-proof p_0 of the last equivalence.

i = 1: $\{h\}(d,1)$ prolongs p_0 to a PA-proof p_1 of
$Prv_{PA}(\ulcorner d \urcorner"\bar{9}, \ulcorner s \urcorner) \longrightarrow \neg\, s$;
this means that if we denote $p_0 * \{h\}(d,1)$ by p_1, then p_1 is a PA-proof of $Prv_{PA}(\ulcorner d \urcorner"\bar{9}, \ulcorner s \urcorner) \longrightarrow \neg\, s$. The last implication will be denoted ANT; it will be used in the antecedent of an instance of A2.

$i = 2$ (First filling): h searches for a PA-proof of

$$\mathrm{Prv}_{PA}(\ulcorner d \urcorner^{\bullet}\overline{9}, \ulcorner \mathrm{ANT} \urcorner);$$

if it succeeds then the found proof is $\{h\}(d,2)$; and we put $p_2 = p_1 * \{h\}(d,2)$. If it fails then $\{h\}(d,2)$ is undefined (as well as $\{h\}(d,i)$ for all $i > 2$). (Similarly for other fillings below.)

Convention : The word "proof" (underlined) will have double meaning in the discussion below; at the moment, think of PA-proofs.

$i = 3$ (Second filling): h searches for a proof of

$$\mathrm{Prv}_{PA}(\ulcorner d \urcorner^{\bullet}\overline{9}, \ulcorner \mathrm{ANT} \urcorner) \longrightarrow \neg\, \mathrm{Prv}_{PA}(\ulcorner d \urcorner^{\bullet}\overline{9}, \ulcorner s \urcorner)$$

which is an instance of A2; if successful the found proof is $\{h\}(d,3)$ and we put $p_3 = p_2 * \{h\}(d,3)$.

$i = 4$: $\{h\}(d,4)$ prolongs p_3 to a proof of $\neg\, \mathrm{Prv}_{PA}(\ulcorner d \urcorner^{\bullet}\overline{9}, \ulcorner s \urcorner)$; we put $p_4 = p_3 * \{h\}(d,4)$.

$i = 5$: $\{h\}(d,5)$ prolongs p_4 to a proof of s (using the equivalence in $i = 0$); we put $p_5 = p_4 * \{h\}(d,5)$.

$i = 6$ (Third filling): h searches for a proof of

$$\mathrm{Prv}_{PA}(\ulcorner d \urcorner^{\bullet}\overline{9}, \ulcorner s \urcorner);$$

if successful then $\{h\}(d,6)$ is the found proof and we put $p_6 = p_5 * \{h\}(d,6)$.

$i = 7$: $\{h\}(d,7)$ prolongs p_6 to a proof of $\neg s$; put $p_7 = p_6 * \{h\}(d,7)$

$i = 8$: $\{h\}(d,8)$ prolongs p_7 to a proof of $s \,\&\, \neg s$.

End of definition of h.

Convention. In the sequel, d will denote an arbitrary Turing machine satisfying the equation

$$\{h\}(d,i) \cong \{d\}(i)$$

for each i = 0,1,... 8. (Existence is given by Recursion Theorem.)

Fact 5. For every choice d according to our convention, d"3 exists and is a PA-proof. (There is a first filling.)

Proof. We know that p_1 is a PA-proof of ANT and that $p_1 = \{h\}(d,0) * \{h\}(d,1)$; since $\{d\}(i) \cong \{h\}(d,i)$ for i = 0,... ... 8, d"2 exists and equals p_1. Thus the formula $Prv_{PA}(\ulcorner d \urcorner " \bar{9}, \ulcorner ANT \urcorner)$ is true and therefore PA-provable (being a Σ_1-formula).

Fact 6. If "proof" in the construction means "PA-proof" then for each d, $\{d\}(3)$ is undefined; the instance of A2 in question is not provable in PA. (Thus d"4 does not exist; there is no second filling.)

Proof. This is because if there were a second filling then there would also exist a third filling (see i = 6 in the construction), since p_4 would be a PA-proof of s and p_4 = d"5; thus $Prv_{PA}(\ulcorner d \urcorner " \bar{9}, \ulcorner s \urcorner)$ would be provable. Thus in that case d"9 would exist and would be a PA-proof of a contradiction.

Theorem 4. The interpretation of T and E by $Prv_{PA}(x,y)$ is not sound in PA. (Immediate from Fact 6.)

Fact 7. Axiom A2 is provable in $PA+Con_{PA}$.

Proof. In $(PA+Con_{PA})$ assume $Prv_{PA}(\ulcorner p \urcorner, \ulcorner Prv_{PA}(\ulcorner p \urcorner, \ulcorner s \urcorner) \longrightarrow \longrightarrow \neg s \urcorner)$ and $Prv_{PA}(\ulcorner p \urcorner, \ulcorner s \urcorner)$. Then $Pr_{PA}(\ulcorner Prv_{PA}(\ulcorner p \urcorner, \ulcorner s \urcorner) \urcorner)$, $Pr_{PA}(\ulcorner Prv_{PA}(\ulcorner p \urcorner, \ulcorner s \urcorner) \longrightarrow \neg s \urcorner)$, hence $Pr_{PA}(\ulcorner \neg s \urcorner)$. But $Prv_{PA}(\ulcorner p \urcorner, \ulcorner s \urcorner)$ implies $Pr_{PA}(\ulcorner s \urcorner)$, so we have $Pr(\ulcorner s \& \neg s \urcorner)$, a contradiction.

Theorem 5. The interpretation of T and E by $Prv_{PA}(x,y)$ is sound w.r.t. $PA+Con_{PA}$. (Immediate.)

Remark. Similarly as in § 3, we can try to understand h as a function trying to construct a proof of a contradiction in $PA+Con_{PA}$. For this purpose, we shall understand the word "proof" (underlined) in the construction as "$(PA+Con_{PA})$-proof".

Fact 8. If the construction of h is modified as just said then d"6 exists and is a $(PA+Con_{PA})$-proof, but for each d, $\{d\}(6)$ diverges thus d"7 does not exist. (There is a second filling but there is no third filling.)

Proof. This follows from Fact 7: A $(PA+Con_{PA})$-proof of A2 is a second filling. Therefore there can be no third filling since otherwise the construction would produce a proof of contradiction in $(PA+Con_{PA})$. (Let us mention that in the present situation p_8 is a $(PA+Con_{PA})$-proof and therefore also $\{d\}(n)$, but it is not a PA-proof. In the second filling, the axiom Con_{PA} is used in a substantial way.)

Thus the Truth-Reaction Paradox does not give us any proof of inconsistency in $PA+Con_{PA}$ because, as we have just seen, there is no third filling.

References

[1] V.S. CHERNIAVSKY: On limitations of artificial intelligence, Inf. Systems 5(1980), 121.

[2] H.B. ENDERTON: Elements of recursion theory, Handbook of Mathematical Logic (North-Holland P.C. 1977), 527-566.

[3] S. FEFERMAN: Transfinite recursive progressions of axiomatic theories, Journ. Symb. Log. 27(1962), 259-316.

[4] I.M. HAVEL: The truth-reaction paradox: a probe of limitations of artificial intelligence, Proc. ECAI 82, Orsay 1982.

[5] C. SMORYŃSKI: The incompleteness theorems, Handbook of Math. Logic (North-Holland P.C. 1977), 821-862.

[6] R. SOLOVAY: Provability interpretations of modal logic, Israel Journ. Math. 25(1976), 287-304.

Mathematical Institute, ČSAV, Žitná 25, 11567 Praha 1, Czechoslovakia

(Oblatum 10.6. 1982)