“NATURAL” REPRESENTATIONS AND EXTENSIONS OF GÖDEL’S SECOND THEOREM

KARL–GEORG NIEBERGALL

Abstract. Building on [9], several types of examples for consistent extensions of PA proving their own consistency, employing only what may be regarded as natural formalizations of their consistency-assertions, are presented. I also suggest and discuss explications of “natural formalization” and analyze the results given before in the light of this discussion.

§1. Introduction. In [9], I have presented a method for obtainig extensions of, e.g., Peano Arithmetic (PA), which are consistent, but nonetheless prove their own consistency. For these results, it is of crucial importance that the theories considered are provided with representations that are “natural”. That is, “T proves its own consistency” is explained as “$T \vdash \mathrm{Con}_\tau$”, where “$\mathrm{Con}_\tau$” is defined in the usual way from the usual arithmetized proof-predicate, and τ is a “natural” representation of a “natural” *axiom-set* of T.[1] How to understand “*natural* representation” is, of course, a decisive question. In [9], I made some suggestions on this topic and, moreover, embedded the metamathematical results given there in some broader philosophical perspective.

This paper is a continuation of [9].[2] Its first part (section 2.1) lists several examples of theories which prove their own consistency employing only natural formalizations of consistency-assertions—in short: *theories proving their own consistency naturally*—and methods to obtain such theories. Whereas in that part, the expression “natural representation” is merely *used*, but not defined, section 3 attempts to provide for the missing definition: I propose and discuss in it several explications of “α is a natural representation of A”.

Cordial thanks to the organizers of LC 2001, in particular Prof. Isaacson and Prof. Visser, for inviting me to that conference. I also would like to thank Prof. H. Friedman and Prof. Willard for their comments on the talk, the referees for their critical remarks and suggestions to improve this paper, and M. Barrios and N. Silich.

[1] In his classical [1], Feferman has shown that—in some sense—even PA can prove its own consistency by supplying a representation pa* of an *axiom-set* of PA such that PA $\vdash \mathrm{Con}_{\mathrm{pa}^*}$ (cf. section 3); but $\mathrm{Pr}_{\mathrm{pa}^*}$ will hardly be regarded as a “natural” representation of PA.

[2] As such, it deals more deeply (see especially sections 3 and 4) with some of the issues I had treated only sketchily or left open in [9], but is rather brief with respect to others: this holds, in particular, for the philosophical background, which is elaborated in [9].

Logic Colloquium ’01
Edited by M. Baaz, S. Friedman, and J. Krajíček
Lecture Notes in Logic, 20

The paper closes with some "negative" metamathematical results (section 4): if all conditions for "α is a natural representation of A" taken into consideration are conjoined, it seems to be hard to find theories proving their own consistency naturally.

§2. **Theories proving their own consistency.** For simplicity, only theories S in L[PA] (or definitional extensions of it) which extend $\mathrm{I}\Sigma_1$ will be considered (Robinson Arithmetic (Q) is an exception; see [12], [4]). Their underlying logic is assumed to be classical first-order logic with identity; it is convenient to axiomatize it such that only *modus ponens* is a primitive rule of inference (see [1] for more details). As usual, a theory S is understood as a set of sentences which is deductively closed; i.e., using "$\overline{A}$" for the set of sentences from L[A] derivable from A, S is a theory $\iff S = \overline{S}$. Let me note that, as such, theories need not be recursively enumerable (r.e.) or even, say, arithmetically definable: think at the *theory of* $\mathcal{N}$, i.e., the set of all sentences from L[PA] being true in $\langle \mathbb{N}, S, +, \cdot, 0\rangle$, as an example for a non arithmetical theory. In fact, non r.e. theories will be the main object of investigation in this paper.[3]

If S fails to be r.e., it is not axiomatizable; for lack of a better term, nonetheless even in this case I will say that S has an *axiom-set*.[4] That is,

$$A \text{ is an } \textit{axiom-set} \text{ for } S \iff S = \overline{A}.$$

Let me emphasize here that for sets of formulas to be theories, they have to be closed only under the usual rules of inference of classical first order logic.[5] Thus, even if a theory S is non r.e., it is r.e. in each of its *axiom-sets*. Moreover, if A is an *axiom-set* for S, S is certainly the set of sentences *derivable* (where the meaning of "derivable" has not been stretched) from A.[6]

Some *axiom-sets* will play a distinguished role for the following investigations: Ax(PA), Ax(Q), Ax($\mathrm{I}\Sigma_n$) ($n \geq 1$) are the usually chosen *axiom-sets* of PA, Q, $\mathrm{I}\Sigma_n$, Ax(PA) being recursive, the others being finite. $\mathrm{Tr}_{\Pi^0_k}$ and $\mathrm{Tr}_{\Sigma^0_k}$ are the sets of true Π^0_k-sentences and the set of true Σ^0_k-sentences ($k \geq 1$).[7]

[3] Besides, it will not be assumed that theories need to be arithmetically sound.

[4] Alternative terminological choices are, e.g., "set of postulates" or "set of pseudoaxioms" (cf. [7]). But see also [1], in which the expression "axiom-system" is used where I prefer "*axiom-set*".

[5] In particular, no infinitary rules (like, e.g., the ω-rule) or rules defined by semantical conditions are employed.

[6] There are further questions, some of which are philosophically more important, like: Do non r.e. theories exist? Are they worthy of studying? Can they be used as *means* of investigation? These deserve a close investigation, which I will not even attempt to carry out here, however (personally, I think that whatever the answers may be, those who assert that the analogous questions concerning r.e. theories should be answered differently are in need of providing reasons for their claims).

[7] See [4] and [5] for the theories and the background on recursion theory and the arithmetical hierarchy.

For the topic “Gödel’s incompleteness theorems”, the arithmetization of metamathematical formulas is of importance.[8] Thus, let $A \subseteq \mathbb{N}^k$ and α a k-place arithmetical formula ($k \geq 1$); in general, $\overline{n}$ will be the numeral denoting the natural number n; then

α is a representation of $A :\Longleftrightarrow$

$$\forall n_1, \dots n_k \in \mathbb{N}\ \big(\langle n_1, \dots, n_k\rangle \in A \Longleftrightarrow \mathcal{N} \models \alpha(\overline{n_1}, \dots, \overline{n_k})\big).$$

I assume that the usual “syntactic” metamathematical vocabulary—like “x is a sequence”, “the length (of sequence) x”, “x is the conditional of y and z”, “x is a formula”, “x is a Π^0_n-formula”—is represented by Σ^0_0-formulas—like “$Seq(x)$”, “$lh(x)$”, “$x = y \dot{\to} z$”, “$Fml(x)$”, “$\Pi^0_n(x)$”—in such a way that conditions characteristic for these notions are provable in the theories considered here (with the exception of Q).[9] For a finite set $\{n_1, \dots, n_k\}$ of natural numbers, the formula “$x = \overline{n_1} \vee \dots \vee x = \overline{n_k}$” is its so called *canonical representation*. In particular, I write “$[q](x)$” and “$[\mathrm{i}\sigma_n](x)$” for the canonical representations of Ax(Q) and Ax($\mathrm{I}\Sigma_n$) ($n \geq 1$). Furthermore, I use “LogAx(x)” for the representation of a recursive set of axioms of first-order logic (to be specific, take the axiomatization from [1]), “pa(x)” for the representations of Ax(PA) and “$\mathrm{Tr}_{\Pi^0_k}(x)$” and “$\mathrm{Tr}_{\Sigma^0_k}(x)$” for the (natural) representations of $\mathrm{Tr}_{\Pi^0_k}$ and $\mathrm{Tr}_{\Sigma^0_k}$ ($k \geq 1$) (see [5] for details). With the exception of “$\mathrm{Tr}_{\Pi^0_k}(x)$” and “$\mathrm{Tr}_{\Sigma^0_k}(x)$”, which are Π^0_k and Σ^0_k, all representations are Σ^0_0.

Given a representation τ of an *axiom-set* T, arithmetical formulas representing the relations or metatheoretical formulas *proof in* τ, *provability in* τ and τ *is consistent* are defined as usual:[10]

Definition 1.

$$\mathrm{Proof}_\tau(x, y) :\Longleftrightarrow Seq(x) \wedge y = x_{lh(x)\dot{-}1} \wedge$$
$$\forall v < lh(x)(\mathrm{LogAx}(x_v) \vee \tau(x_v) \vee \exists uw < v(x_w = x_u \dot{\to} x_v)),$$
$$\mathrm{Pr}_\tau(y) :\Longleftrightarrow \exists x\ \mathrm{Proof}_\tau(x, y),$$
$$\mathrm{Con}_\tau :\Longleftrightarrow \neg\,\mathrm{Pr}_\tau(\overline{\ulcorner \bot \urcorner}).$$

Furthermore, RFN[σ] is the uniform reflection principle for S (σ, to be more precise), i.e., the set of all formulas “$\forall x(\mathrm{Pr}_\sigma(\overline{\ulcorner \psi(\dot{x})\urcorner}) \to \psi(x))$” for arbitrary formulas ψ in L[PA]; $\mathrm{RFN}_{\Sigma^0_k}[\sigma]$ is the restriction of RFN[σ] to the

[8] With respect to the notation employed, I follow [1]; cf. also [11]; in particular, I employ the *dot-notation* presented there. For questions of provability in theories weaker than PA, see [4].

[9] Some gödelization is presupposed here: for an expression t from L[PA] (or some if its extensions), $\ulcorner t \urcorner$ is the Gödel-number and $\overline{\ulcorner t \urcorner}$ the Gödel-numeral of t. Given this, α is called a representation of Σ, too, if $A = \{\ulcorner \psi \urcorner \mid \psi \in \Sigma\}$ and α represents A.

[10] Usually, the phrases “proof in T”, “provability in T” and “T is consistent” are used here. Besides, one should perhaps say “derivation” and “derivable” instead of “proof” and “provable”; but the use of the latter is more common in metamathematics.

Σ^0_k-formulas; and Rfn[σ] is the local reflection principle for S, i.e., the set of all sentences "$\mathrm{Pr}_\sigma(\overline{\ulcorner\psi\urcorner}) \to \psi$" for arbitrary sentences ψ in L[PA].

In this context,[11] Gödel's Second Incompleteness Theorem can be stated quite generally as follows:

Gödel's Second Theorem. For each consistent r.e. extension T of PA and for each representation τ of an *axiom-set* of T which is Σ^0_1, $T \nvdash \mathrm{Con}_\tau$.

Thus, theories extending PA which prove their own consistency naturally can probably not be found under the r.e. ones (if the interpretation of "naturally" is not stretched too much).

2.1. Mutual consistency proofs. The examples of theories proving their own consistency naturally given in [9] rest on the fact (taken from [8]) that there exist theories S, T such that S proves the consistency of T and T proves the consistency of S (while Gödel's incompleteness theorems hold for both).

First example for mutual consistency proofs:

Let $S_1 := \mathrm{PA} + \mathrm{Tr}_{\Pi^0_1}$. S_1 has the natural *axiom-set* $\mathrm{Ax}(\mathrm{PA}) \cup \mathrm{Tr}_{\Pi^0_1}$, which is represented in a natural way by the formula σ_1, where

$$\sigma_1(x) :\longleftrightarrow \mathsf{pa}(x) \vee \mathrm{Tr}_{\Pi^0_1}(x).$$

Now, let $T_1 := \mathrm{PA} + \mathrm{Con}_{\sigma_1}$ (i.e., $T_1 := \mathrm{PA} + \mathrm{Con}_{\mathsf{pa}+\mathrm{Tr}_{\Pi^0_1}}$). T_1 has the natural *axiom-set* $\mathrm{Ax}(\mathrm{PA}) \cup \{\mathrm{Con}_{\sigma_1}\}$, which is represented in a natural way by the formula τ_1, where

$$\tau_1(x) :\longleftrightarrow \mathsf{pa}(x) \vee x = \overline{\ulcorner \mathrm{Con}_{\sigma_1} \urcorner}.$$

By definition of T_1, $T_1 \vdash \mathrm{Con}_{\sigma_1}$. And since T_1 is r.e. and consistent, its consistency assertion Con_{τ_1} is a true Π^0_1-sentence, whence provable in S_1.

Second example for mutual consistency proofs:

Let $S_{2,k} := \mathrm{I}\Sigma_1 + \mathrm{Tr}_{\Pi^0_{k+1}}$ ($k \in \mathbb{N}$), $T_2 := \mathrm{PA}$. $S_{2,k}$ and T_2 have natural *axiom-sets* $\mathrm{Ax}(\mathrm{I}\Sigma_1) \cup \mathrm{Tr}_{\Pi^0_{k+1}}$ and $\mathrm{Ax}(\mathrm{PA})$. These *axiom-sets* are represented in a natural way by the formulas $\sigma_{2,k}$ and τ_2, where

$$\sigma_{2,k}(x) :\longleftrightarrow [\mathsf{i}\sigma_1](x) \vee \mathrm{Tr}_{\Pi^0_{k+1}}(x), \quad \text{and} \quad \tau_2(x) :\longleftrightarrow \mathsf{pa}(x).$$

In [8], it is shown that $T_2 \vdash \mathrm{RFN}[\sigma_{2,k}]$ for each $k \in \mathbb{N}$.[12] Since $\mathrm{RFN}_{\Sigma^0_k}[\tau_2]$ is a set of true Π^0_{k+1}-sentences, $S_{2,k} \vdash \mathrm{RFN}_{\Sigma^0_k}[\tau_2]$.

[11] It may be that very weak theories formulated in L[PA] or languages with similar or less expressive richness prove their own consistency; see [14] for work going in this direction and further remarks on the relevant literature.

[12] Provability in $S_{2,k}$ is usually arithmetized in a way different from $\mathrm{Pr}_{\sigma_{2,k}}$ (see e.g., [11]), making it easier to formalize metamathematical arguments about $S_{2,k}$. But $\sigma_{2,k}$ is surely the representation which is intuitively more natural.

Thus, it is not only possible to obtain mutual consistency proofs, but $S_{2,k}$ and T_2 provide for examples of the mutual provability of partial soundness (up to an arbitrarily given extent).[13]

2.2. System-internal consistency proofs. Having theories proving mutually their consistency at hand, it is easy to obtain theories proving their own consistency (naturally): one simply has to take intersections of theories proving their mutual consistency.

In order to carry out this idea, one should have control over representations of intersections of theories. Thus, let S and T be arithmetically definable theories and $\sigma(x)$ and $\tau(x)$ be representations of *axiom-sets* of S and T. Then "$\mathrm{Pr}_\sigma(x)$" and "$\mathrm{Pr}_\tau(x)$" are representations of the theories S and T, and "$\mathrm{Pr}_\sigma(x) \wedge \mathrm{Pr}_\tau(x)$" is a representation of $S \cap T$. Moreover, it seems plausible to me that this formula is a *natural* representation of $S \cap T$, if the representations σ and τ are natural (see section 3 for more on this). Thus, I define

$$(\sigma \wedge \tau)(x) :\Longleftrightarrow \mathrm{Pr}_\sigma(x) \wedge \mathrm{Pr}_\tau(x).$$

Now, by what has been said so far, it also seems that "$\mathrm{Pr}_{\sigma\wedge\tau}$" is a natural representation of $\overline{S \cap T}$. But $\overline{S \cap T} = S \cap T$—whence we have two "natural" representations ("$\sigma \wedge \tau$" and "$\mathrm{Pr}_{\sigma\wedge\tau}$") for just one object (the theory $S \cap T$). In general, I do not think that the mere circumstance that we have several representations of the same theory presents a problem (see again section 3). Besides, the case considered here seems to me particularly harmless: the reason is that these two representations are PA-provably equivalent.

Lemma 1. (a) $\mathrm{PA} \vdash \forall x\, ((\sigma \wedge \tau)(x) \longleftrightarrow \mathrm{Pr}_{\sigma\wedge\tau}(x))$.
(b) $\mathrm{PA} \vdash \mathrm{Con}_\sigma \vee \mathrm{Con}_\tau \longleftrightarrow \mathrm{Con}_{\sigma\wedge\tau}$.

Example 1. Take S_1, T_1 and σ_1, τ_1 from the first example.

Since $S_1 \vdash \mathrm{Con}_{\tau_1}$, it follows that $S_1 \vdash \mathrm{Con}_{\sigma_1\wedge\tau_1}$.
Since $T_1 \vdash \mathrm{Con}_{\sigma_1}$, it follows that $T_1 \vdash \mathrm{Con}_{\sigma_1\wedge\tau_1}$.

Therefore, $S_1 \cap T_1 \vdash \mathrm{Con}_{\sigma_1\wedge\tau_1}$.

Example 2. Take $S_{2,k}, T_2$ and $\sigma_{2,k}, \tau_2$ ($k \in \mathbb{N}$) from the second example.

Since $T_2 \vdash \mathrm{RFN}[\sigma_{2,k}]$, it follows that $T_2 \vdash \mathrm{RFN}[\sigma_{2,k} \wedge \tau_2]$.
Since $S_{2,k} \vdash \mathrm{RFN}_{\Sigma^0_k}[\tau_2]$, it follows that $S_{2,k} \vdash \mathrm{RFN}_{\Sigma^0_k}[\sigma_{2,k} \wedge \tau_2]$.

Therefore, $S_{2,k} \cap T_2 \vdash \mathrm{RFN}_{\Sigma^0_k}[\sigma_{2,k} \wedge \tau_2]$.

Thus, consistent extensions of PA can also prove their own soundness naturally up to an arbitrarily preassigned degree.

The foregoing considerations can be found (more or less) already in [9]. Coming to the new work in what follows, let me first give a further example.

[13] Evidently, the second example is obtained as a special case of a general procedure for building theories which mutually prove their consistency. In fact, there are many theories which mutually prove their consistency or partial soundness.

Example 3. Let $(\mathrm{PA}_n)_{n\in\mathbb{N}}$ be a sequence (as discovered by Feferman, Friedman, Solovay and possibly others; see [11]) of consistent r.e. extensions of PA such $\forall n \in \mathbb{N}\, \mathrm{PA}_n \vdash \mathrm{Con}_{\mathrm{pa}_{n+1}}$. To be more specific, I follow the version presented in [6].

Here (for each $n \in \mathbb{N}$), $\mathrm{PA}_n := \mathrm{PA} + \beta(\overline{n})$, where each $\beta(\overline{n})$ is a certain Π^0_1-sentence satisfying

$$\mathrm{PA}_n \vdash \beta(\overline{n+1}) \wedge \mathrm{Con}_{\mathrm{pa}+\beta(\overline{n+1})}\,.^{14}$$

Now define $(S_3 :=)\ \mathrm{PA}_\omega := \bigcap\{\mathrm{PA}_n \mid n \in \mathbb{N}\}$, and

$$\alpha(x,y) :\longleftrightarrow \mathrm{pa}(x) \vee x = \overline{\ulcorner \beta(\dot{y}) \urcorner},$$
$$\mathrm{pa}_\omega(x) :\longleftrightarrow \forall y\ \mathrm{Pr}_{\alpha(,y)}(x).^{15}$$

Then,

$$\begin{aligned}\mathcal{N} \models \alpha(\overline{\ulcorner\psi\urcorner}, \overline{n}) &\iff \mathcal{N} \models \mathrm{pa}(\overline{\ulcorner\psi\urcorner}) \vee \overline{\ulcorner\psi\urcorner} = \overline{\ulcorner\beta(\overline{n})\urcorner} \\ &\iff \psi \in \mathrm{Ax}(\mathrm{PA}) \cup \{\beta(\overline{n})\},\end{aligned}$$

and

$$\mathcal{N} \models \mathrm{pa}_\omega(\overline{\ulcorner\psi\urcorner}) \iff \forall n\ (\mathrm{Ax}(\mathrm{PA})\ \cup \{\beta(\overline{n})\}) \vdash \psi \iff \mathrm{PA}_\omega \vdash \psi.$$

That is, pa_ω is a—I think, natural—representation of the theory PA_ω. But there is, like before, a further representation of the latter which deserves to be called "natural": $\mathrm{Pr}_{\mathrm{pa}_\omega}$. Yet, again, these two representations are PA-provably equivalent with each other:

Claim. $\mathrm{PA} \vdash \forall x\ (\mathrm{pa}_\omega(x) \longleftrightarrow \mathrm{Pr}_{\mathrm{pa}_\omega}(x))$.

Proof. Only "$\longleftarrow$" has to be proved. But since $\mathrm{PA} \vdash \forall xy\ (\mathrm{pa}_\omega(x) \wedge \mathrm{pa}_\omega(x \mathbin{\dot{\rightarrow}} y) \longrightarrow \mathrm{pa}_\omega(y))$, this direction clearly holds, too (cf. [1]).

Claim. $\mathrm{PA}_\omega \vdash \mathrm{Con}_{\mathrm{pa}_\omega}$.

Proof. Surely, for each $n \in \mathbb{N}$, $\mathrm{PA} \vdash \forall x\ (\forall y\ \mathrm{Pr}_{\alpha(,y)}(x) \longrightarrow \mathrm{Pr}_{\mathrm{pa}+\beta(\overline{n})}(x))$, i.e.,

$$\mathrm{PA} \vdash \forall x\ (\mathrm{pa}_\omega(x) \longrightarrow \mathrm{Pr}_{\mathrm{pa}_n}(x)),$$

whence (by the first claim) $\mathrm{PA} \vdash \mathrm{Con}_{\mathrm{pa}_n} \longrightarrow \mathrm{Con}_{\mathrm{pa}_\omega}$.

Since $\mathrm{PA}_n \vdash \mathrm{Con}_{\mathrm{pa}_{n+1}}$, $\mathrm{PA}_n \vdash \mathrm{Con}_{\mathrm{pa}_\omega}$ is obtained. As n was arbitrary, it follows that $\mathrm{PA}_\omega \vdash \mathrm{Con}_{\mathrm{pa}_\omega}$.

As a next step, let me present some methods of proceeding from theories proving their own consistency to other types of theories which also prove their own consistency. For this, let σ, τ be representations of *axiom-sets* of consistent arithmetically definable theories S, T extending $\mathrm{I}\Sigma_1$.

[14] Therefore, each PA_n is a sound r.e. extension of $\mathrm{PA} + \mathrm{Con}_{\mathrm{pa}}$.

[15] This notation is taken from [2].

Method 1. Add a consistent Σ^0_1-sentence.

Lemma 2. If $S \vdash \mathrm{Con}_\sigma$ and ψ is Σ^0_1, and $S \vdash \forall x\ (\mathrm{Pr}_\mathrm{q}(x) \to \mathrm{Pr}_\sigma(x))$,[16] then $S + \psi \vdash \mathrm{Con}_{\sigma+\psi}$.

Proof. Since $S \vdash \forall x\ (\mathrm{Pr}_\mathrm{q}(x) \to \mathrm{Pr}_\sigma(x))$ and Q is S-provably Σ^0_1-complete (S being an extension of $\mathrm{I}\Sigma_1$), Con_σ implies $\mathrm{Rfn}_{\Pi^0_1}[\sigma]$ over S. Therefore, $S \vdash \mathrm{Pr}_{\sigma+\psi}(\overline{\ulcorner \bot \urcorner}) \longrightarrow \mathrm{Pr}_\sigma(\overline{\ulcorner \psi \to\bot \urcorner}) \longrightarrow (\psi \to\bot)$, for $S \vdash \mathrm{Con}_\sigma$ and $\psi \to\bot$ is Π^0_1. It follows that $S + \psi \vdash \neg\, \mathrm{Pr}_{\sigma+\psi}(\overline{\ulcorner \bot \urcorner})$.

Method 2. Intersect with a theory proving its own consistency.

Lemma 3. If $S \vdash \mathrm{Con}_\sigma$ and $T \vdash \mathrm{Con}_\sigma$, then $S \cap T \vdash \mathrm{Con}_{\sigma\wedge\tau}$.

Proof. Since $S \cap T \vdash \mathrm{Con}_\sigma$, the conclusion follows by Lemma 1.

Method 3. Maximize in a provable way.

Each consistent arithmetically definable theory has arithmetically definable completions, i.e., maximally consistent extensions; moreover, this statement can be formalized in $\mathrm{I}\Sigma_1$ (see [1], [6] and especially [13] for more detailed presentations of this "formalized completeness theorem"). A weakened version of this metatheorem which suffices here is:

For each representation σ of some *axiom-set* there is an arithmetical formula $v(\sigma)$ representing an *axiom-set* of a complete extension $V(S)$ of S such that

$$\mathrm{I}\Sigma_1 + \mathrm{Con}_\sigma \vdash \mathrm{Con}_{v(\sigma)}\,.$$

In addition, $v(\sigma)$ can be found in Δ^0_{k+1} if σ is Σ^0_k $(k \geq 1)$.

Now, let S be an extension of $\mathrm{I}\Sigma_1$ with representation σ such that $S \vdash \mathrm{Con}_\sigma$. Then $V(S) \vdash \mathrm{Con}_{v(\sigma)}$; i.e., $V(S)$ is a complete extension of PA proving its own consistency.

Method 4. Intersect provable extensions.

Lemma 4. Assume $S \cap T \vdash \mathrm{Con}_{\sigma\wedge\tau}$; and let α, β be arithmetical formulas such that $S \cap T \vdash \mathrm{Con}_\sigma \to \mathrm{Con}_\alpha$ and $S \cap T \vdash \mathrm{Con}_\tau \to \mathrm{Con}_\beta$. Then $S \cap T \vdash \mathrm{Con}_{\alpha\wedge\beta}$.

Proof. Since $\mathrm{PA} \vdash \mathrm{Con}_{\sigma\wedge\tau} \longrightarrow \mathrm{Con}_\sigma \vee \mathrm{Con}_\tau$ and $\mathrm{PA} \vdash \mathrm{Con}_\alpha \vee \mathrm{Con}_\beta \longrightarrow \mathrm{Con}_{\alpha\wedge\beta}$ (by Lemma 1), the assumptions yield the claim.

I give some applications of these methods.

Example 4 (by method 1). Let $S_4 := \mathrm{PA}_\omega + \{\neg\, \mathrm{Con}_{\mathrm{pa}_1}\}$, $\sigma_4(x) := \mathrm{pa}_\omega(x) \vee x = \overline{\ulcorner \neg\, \mathrm{Con}_{\mathrm{pa}_1} \urcorner}$. S_4 is consistent, and $S_4 \vdash \mathrm{Con}_{\sigma_4}$ by Lemma 2.

For the following examples, let $T_5 := S_{2,0} \cap T_2$ with representation $\tau_5(x) :\longleftrightarrow \sigma_{2,0}(x) \wedge \tau_2(x)$.

[16] By adding "$[q](x)$" to "$\sigma(x)$", this can always be achieved.

Example 5 (by method 2). Take $\mathrm{I}\Sigma_1 + \mathrm{Con}_{\tau_5}$ with natural representation "$[\mathrm{i}\sigma_1](x) \vee x = \overline{\ulcorner \mathrm{Con}_{\tau_5} \urcorner}$". Since $T_5 \vdash \mathrm{Con}_{\tau_5}$, by method 2, $(\mathrm{I}\Sigma_1 + \mathrm{Con}_{\tau_5}) \cap T_5$ proves its own consistency with representation $([\mathrm{i}\sigma_1] + \mathrm{Con}_{\tau_5}) \wedge \tau_5$.

Example 6 (by method 3). Since $T_5 \vdash \mathrm{Con}_{\tau_5}$, there is a completion $V(T_5)$ with representation $v(\tau_5)$ in Δ^0_3 such that $V(T_5) \vdash \mathrm{Con}_{v(\tau_5)}$.

Example 7 (by method 4). Let $V(S_{2,0})$ be a completion in Δ^0_3 of $S_{2,0}$ with an *axiom-set* represented by $v(\sigma_{2,0})$ and $V(T_2)$ be a completion in Δ^0_2 of T_2 with an *axiom-set* represented by $v(\tau_2)$ such that

$$T_5 + \mathrm{Con}_{\sigma_{2,0}} \vdash \mathrm{Con}_{v(\sigma_{2,0})} \text{ and } T_5 + \mathrm{Con}_{\tau_2} \vdash \mathrm{Con}_{v(\tau_2)}\,.$$

Then $T_5 \vdash \mathrm{Con}_{v(\sigma_{2,0}) \wedge v(\tau_2)}$ by Lemma 4, and $V(S_{2,0}) \cap V(T_2)$ proves its own consistency with representation $v(\sigma_{2,0}) \wedge v(\tau_2)$.

§3. **Natural representations.** In the previous section, I have presented some methods which allow the construction of many theories which prove their own consistency. Moreover, I have repeatedly asserted that the representations of those theories considered there are natural, but not supplied any arguments for these claims. In fact, without a conceptual analysis of "natural representation", it seems hardly possible to do so. In this section, I will attempt to close that gap.[17]

When it comes to the task of formulating an explication of "α is a natural representation of A", there is no sense in denying that we share linguistic intuitions about which representations are natural and which ones fail to be natural.—On the one hand, there should be "universal" agreement on several paradigmatic cases—where we could just *stipulate* which representations are natural and unnatural (or artificial). Here, we should find (or so do I believe) the usual representations "pa(x)" of Ax(PA) and the the usual representations of the recursive syntactic notions—like "$x = y \dot{\rightarrow} z$", "$\mathit{Fml}(x)$", "$\Pi^0_n(x)$"—but also such formulas as "$x + y = z$" for the addition function. Furthermore, finite sets of formulas $\{\psi_1, \ldots, \psi_k\}$ are naturally represented by their *canonical representation* (see section 2.1).—On the other hand, we have clear intuitions about how to obtain natural representations of B from natural representations of $A_1, \ldots, A_n$, provided B is constructed by simple methods from $A_1, \ldots, A_n$. Let me give two examples:

(a) if α is a natural representation of A, then "Pr_α" is a natural representation of $\overline{A}$;

(b) if α is a natural representation of A and ψ is a sentence, then "$\alpha(x) \vee x = \overline{\ulcorner \psi \urcorner}$" is a natural representation of $A \cup \{\psi\}$.

What I will certainly not assume is that each set of natural numbers (e.g., Gödel numbers) has only one natural representation. Quite the opposite: I

[17] For further considerations on this topic, see [9].

think each one has infinitely many natural representations.[18] If, e.g., $\alpha(x)$ is a natural representation of A, what about "$\alpha(x) \wedge x = x$"? More interesting is the following example: for each theorem ψ of A, "$\mathrm{Pr}_{\alpha+\psi}(x)$" is a natural representation of $\overline{A+\psi}$ (by (a) and (b)); but this set is just $\overline{A}$ (but see section 3.4).[19]

Surely, these examples and considerations do not automatically deliver a precise and generally adequate explication of "α is a natural representation of A". But they also should make it plausible that the search for such an explication is not futile. Thus, in the remaining part of this section, I will propose three ways of giving sufficient or necessary conditions for "α is a natural representation of A" which should be faithful to what I said in these introductory reflections.[20]

3.1. The inductive approach. The idea that natural representations are built from natural representations by using procedures which lead from natural to natural representations is an *inductive conception of naturality*. Rather than using it as a mere means for testing the adequacy of definitions of "α is a natural representation of A", I will employ it to *obtain* such a definition.[21]

Base.

(B1) (by stipulation[22]): "$\mathrm{pa}(x)$" is a natural representation of $\mathrm{Ax}(\mathrm{PA})$ and "$[\mathrm{i}\sigma_1](x)$" is a natural representation of $\mathrm{Ax}(\mathrm{I}\,\Sigma_1)$. "$\Pi^0_k(x)$" and "$\Sigma^0_k(x)$" (see [5]) are natural representations of the sets (of Gödel numbers) of Π^0_k- and Σ^0_k-formulas ($k \geq 1$). The usual function signs for some primitive recursive functions (like successor, addition, multiplication) are natural representations of those functions.

(B2) (by a general principle): if Σ is a finite set of formulas $\psi_1, \dots, \psi_n$ and $A_\Sigma = \{\ulcorner\psi\urcorner \mid \psi \in \Sigma\}$, then

$$x = \overline{\ulcorner\psi_1\urcorner} \vee \cdots \vee x = \overline{\ulcorner\psi_n\urcorner}$$

is a natural representation of A_Σ (and Σ).

[18] I agree that it would also be interesting to have a more fine-grained conception of natural representation.

[19] Sometimes, different axiom systems are common for one theory: Zermelo-Fraenkel set theory (ZF) is an interesting example, with $\mathrm{Ax}_1(\mathrm{ZF})$ being its recursive *axiom-set* containing the axiom scheme of replacement and $\mathrm{Ax}_2(\mathrm{ZF})$ being its recursive *axiom-set* containing the axiom scheme of collection. This leads to two natural representations, "$\mathrm{zf}_1(x)$" and "$\mathrm{zf}_2(x)$" of these *axiom-sets* and to two natural representations of ZF itself.

[20] It may be asked "What about Löb's derivibality conditions for theories which prove their own consistency?" Provable closure under *modus ponens* is, of course, for free. Moreover, not all of the derivability conditions can hold for such theories, because this would imply the unprovability of consistency for them. Of the two remaining ones, closure under necessitation is the least one could hope for. See section 4 for more on this.

[21] There is nothing new to the following clauses; it just seems that they have not been put to work as part of an explication of "α is a natural representation of A".

[22] The list given here is certainly not supposed to be exhaustive. I think this has to remain like this: for one, it is open which theories and, therefore, representations could one day be taken to be worthy of investigation.

Step. Let $\alpha(x)$, $\beta(x)$ be a $k+1$-place arithmetical formulas and $A, B \subseteq \mathbb{N}^{k+1}$; then

(I1) If $\alpha(x)$ is a natural representation of A and $\beta(x)$ is a natural representation of B, then $\alpha(x) \vee \beta(x)$ is a natural representation of $A \cup B$;
(I2) If $\alpha(x)$ is a natural representation of A and $\beta(x)$ is a natural representation of B, then $\alpha(x) \wedge \beta(x)$ is a natural representation of $A \cap B$;
(I3) If $\alpha(x)$ is a natural representation of A, then $\neg\alpha(x)$ is a natural representation of $\mathbb{N}^{k+1} \setminus A$;
(I4) If $\alpha(x)$ is a natural representation of A, then $\mathrm{Pr}_\alpha(x)$ is a natural representation of $\overline{A}$;
(I5) If for each $n \in \mathbb{N}$, $\alpha(x, \overline{n})$ is a natural representation of $\{k \in \mathbb{N} \mid \langle k, n\rangle \in A\}$, then $\forall y\ \alpha(x, y)$ is a natural representation of $\{k \in \mathbb{N} \mid \forall n\ (\langle k, n\rangle \in A)\}$.

Surely, if no clauses are added to the ones given here for the inductive definition of "α is a natural representation of A", then α is a representation of A if it is a natural representation of A. But this set of conditions is supposed to be merely a first suggestion: it may not contain *all* relevant conditions.

3.2. The complexity approach. When a theory T is r.e. (that is, Σ_1^0), it does not only have a r.e. set of axioms (which is trivial), but even a recursive one (Craig's theorem). These sets of axioms can be represented by a Σ_1^0- and a Σ_0^0-formula, resp. An analogous result holds if T is Σ_{k+1}^0 ($k \in \mathbb{N}$): in this case, T has an *axiom-set* of complexity Π_k^0 (see [3]). Of course, a r.e. theory T may also have an *axiom-set* A_T which has a higher complexity than Σ_1^0. But it seems that such an *axiom-set* must be quite unusual, if not odd.

This suggests that for a representation α of A to be natural, it should not be unnecessarily complex. More formally put, let α be a k-place arithmetical formula and $B \subset \mathbb{N}^k$ be arithmetically definable. In comparing the complexity of these objects, I will take as relevant only what is expressible in the arithmetical hierachy. Thus, with "x" and "y" standing for "α" or "B", I define:

$x <_{Co} y :\Longleftrightarrow$ for some $n \in \mathbb{N}$, x is Σ_n^0 and y fails to be Σ_n^0, or x is Δ_n^0 and y fails to be Δ_n^0.[23]

My second proposal for (a necessary condition for) "α is a natural representation of A" is now as follows (for arithmetical α representing A):

If α is a natural representation of A, then $\alpha \leq_{Co} \overline{A}$.[24]

[23] For a theory T, let's explain "α is $\Sigma_n^{0,T}$ / $\Pi_n^{0,T}$ / $\Delta_n^{0,T}$" as "There is a Σ_n^0-formula φ / Π_n^0-formula ψ / Σ_n^0-formula χ_1 and Π_n^0-formula χ_2 such that $T \vdash \alpha \leftrightarrow \varphi$ / $T \vdash \alpha \leftrightarrow \psi$ / $T \vdash \alpha \leftrightarrow \chi_1$ and $T \vdash \alpha \leftrightarrow \chi_2$" and deal similarly with A instead of α. Now, since it may be preferable to regard bounded quantifiers as not introducing further complexity, one should opt rather for the following definition: "$x <_{Co} y :\Longleftrightarrow$ for some $n \in \mathbb{N}$, x is $\Sigma_n^{0,\mathrm{PA}}$ and y fails to be $\Sigma_n^{0,\mathrm{PA}}$, or x is $\Delta_n^{0,\mathrm{PA}}$ and y fails to be $\Delta_n^{0,\mathrm{PA}}$."

[24] "$x \leq_{Co} y$" is defined as "$\neg(y <_{Co} x)$".

3.3. The numeration approach. If τ is a representation of (an *axiom-set* of) T, but T “does not notice it”, then “strange” results about T-provable sentences in which τ occurs may not be as strange after all. For example, τ may be chosen in such a way that, in T, “Pr_τ” behaves like “$\mathrm{Pr}_{[\mathrm{E}]}$” for some finitely axiomatizable subtheory E of T. If, now, T is reflexive, it may prove “Con_τ”; but this is nothing to wonder about.

The anthropomorphic “T does not notice that τ is a representation of (an *axiom-set*) of T” may be explained in two ways. One is: there are certain conditions which actually hold for formulas containing τ, but it is not T-provable that they do.[25] The second is: τ does not numerate (an *axiom-set*) of T in T. Here,

$$\alpha \text{ numerates } A \text{ in } S :\Longleftrightarrow$$
$$\forall n_1, \dots n_k \in \mathbb{N}\big(\langle n_1, \dots, n_k\rangle \in A \Longleftrightarrow S \vdash \alpha(\overline{n_1}, \dots, \overline{n_k})\big).$$

I will only deal with the second idea. This, then, is my third suggestion for (a necessary condition for) “α is a natural representation of A” (for arithmetical α representing A):

If α is a natural representation of A, then α numerates A in $\overline{A}$.

3.4. Comments on the approaches: positive instances and open ends. To begin with, let’s have a look at the examples presented in section 2 from the perspective of the definitions given in section 3.

Relative to the *inductive approach*—and apart from the cases where complete theories are taken into account[26]—all representations both of the theories and their *axiom-sets* discussed above are natural. Let me show this for PA_ω and pa_ω:

By the base clauses, “$\mathrm{pa}(x)$” and “$x = \overline{\ulcorner\beta(\overline{n})\urcorner}$” are natural representation of $\mathrm{Ax}(\mathrm{PA})$ and $\{\beta(\overline{n})\}$. Thus, by the induction step for “$\vee$”, “$\alpha(x, \overline{n})$” is a natural representation of $\mathrm{Ax}(\mathrm{PA}) \cup \{\beta(\overline{n})\}$ for each $n \in \mathbb{N}$. Therefore, for each $n \in \mathbb{N}$, “$\mathrm{Pr}_{\alpha(,\overline{n})}(x)$” is a natural representation of $\{\psi \mid \mathrm{Ax}(\mathrm{PA}) \cup \{\beta(\overline{n})\} \vdash \psi\}$. Finally, by the induction step for the universal quantifier, “$\mathrm{pa}_\omega(x)$” is a natural representation of $\{\psi \mid \forall n\ (\mathrm{Ax}(\mathrm{PA}) \cup \{\beta(\overline{n})\}) \vdash \psi\}$, i.e., of PA_ω.

Although these considerations are an indication that the set of natural representations of arithmetically definable theories is wide enough, there are examples which suggest that this may not be so: there just seems to be too much dependency on notational features built into it. Take, e.g., the formula “$\neg\forall y \neg\, \mathrm{Proof}_{\mathrm{pa}}(y, x)$”; the inductive analysis does not put it into the set of natural representations of PA—it simply does not have the right form. Yet,

[25] This has to do with a certain understanding of “intensional” to be found in proof-theory; see [1].

[26] The inductive definition of “α is a natural representation of A” does not contain any clause dealing with complete theories. At the moment, I do not have strong intuitions on how to include one.

the fact that this formula is provably equivalent to "$\mathrm{Pr}_{\mathrm{pa}}(x)$" may motivate putting the following principle on top of the inductive definition:

- If $\alpha(x)$ is a natural representation of A and $\vdash \forall x\ (\alpha(x) \leftrightarrow \beta(x))$, then $\beta(x)$ is a natural representation of A.

But perhaps logical equivalence is too narrow. In fact, consider the formula pa'_ω defined by

$$\mathrm{pa}'_\omega(x) :\longleftrightarrow \forall y\ (\mathrm{Pr}_{\mathrm{pa}}(\overline{\ulcorner\beta(\dot{y})\urcorner}\dot{\to} x)).$$

pa'_ω is a representation of PA_ω which seems to be as natural as pa_ω; and, again, it is not declared to be a natural representation of PA_ω by the inductive definition. But now, "$\forall x\ (\mathrm{pa}_\omega(x) \longleftrightarrow \mathrm{pa}'_\omega(x))$" is not provable in first-order predicate logic alone. As a reply, the previous extension of the inductive approach could be strengthened to

- If $\alpha(x)$ is a natural representation of A and $T \vdash \forall x\ (\alpha(x) \leftrightarrow \beta(x))$, then $\beta(x)$ is a natural representation of A.

The problem with this principle, however, is its dependency on a theory T: which one should we choose? T should not be too weak for its purpose of obtaining sufficiently many "new" natural representations of A. From this viewpoint, PA is a good choice.[27] Or should we take some theory α is about for T (i.e., PA_ω in our example)? Furthermore, it would be possible to leave T just as it is—as an additional parameter. Yet, in this case we would get a notion of *T-natural representation* instead of the originally sought *representation*.—Nevertheless, I think that *some* version of the principle should be added to the inductive definition.[28]

In case of the *complexity approach*, I consider the situation as somewhat problematic in the sense that for most of the *theories* dealt with in section 2—i.e., for $S_1 \cap T_1$, $S_{2,k} \cap T_2$, S_3, S_4 and $V(T_5)$—their arithmetical complexity is not known to me.[29] In example 5, we have a theory with an *axiom-set* which is Σ^0_2; but that theory is just $\mathrm{I}\Sigma_1 + \mathrm{Con}_{\tau_5}$, which is r.e.. Thus, the representation "$([\mathrm{i}\sigma_1] + \mathrm{Con}_{\tau_5}) \wedge \tau_5$" is not a natural one.

[27] See also [9]; in fact, PA proves "$\forall x\ (\mathrm{pa}_\omega(x) \longleftrightarrow \mathrm{pa}'_\omega(x))$".

[28] As sort of a converse, it may be plausible to conjecture that if Pr_τ and $\mathrm{Pr}_{\tau'}$ are natural representations of one theory T, they are T- (or PA-) provably equivalent. But I think there are counterexamples: take, e.g., $S_{2,0}$. One *axiom-set* of it is $\mathrm{Ax(PA)} \cup \mathrm{Tr}_{\Pi^0_1}$. Since $S_{2,0}$ proves (each instance of the scheme of) *local ω-consistency* for PA [10] (in short: ω-Con[pa]), $\mathrm{Ax(PA)} \cup \mathrm{Tr}_{\Pi^0_1} \cup \omega$-Con[pa] is an *axiom-set* for $S_{2,0}$, too. Now, take a natural representation $\omega Con_{\mathrm{pa}}(x)$ of ω-Con[pa] and the natural representations $\sigma_{2,0}$ and $\sigma_{2,0}(x) \vee \omega Con_{\mathrm{pa}}(x)$ of the *axiom-sets*. It turns out that not even $S_{2,0}$ can prove "$\forall x\ (\mathrm{Pr}_{\sigma_{2,0}}(x) \leftrightarrow \mathrm{Pr}_{\sigma_{2,0}+\omega Con_{\mathrm{pa}}}(x))$"; see [8].

[29] If, e.g., PA_ω is Π^0_2, the complexity condition is satisfied and pa_ω may well be taken to be a natural representation of PA_ω. But if this theory is r.e., it's representation pa_ω is distinguished from Feferman's pa* only in that *intuitively* (what I would still claim to hold), pa_ω is a natural representation of PA_ω, whereas $\mathrm{Pr}_{\mathrm{pa}*}$ is no natural representation of PA.

Coming to example 7, the theory $V(S_{2,0}) \cap V(T_2)$ is Δ^0_3 (being an intersection of a Δ^0_3- and a Δ^0_2-set). Moreover, it has no lower complexity: for assume it were Σ^0_2; since $V(S_{2,0})$ and $V(T_2)$ are both complete, they are inconsistent with each other; thus, for some sentence φ, $(V(S_{2,0}) \cap V(T_2)) + \varphi = V(S_{2,0})$. Yet, following from the assumption, $(V(S_{2,0}) \cap V(T_2)) + \varphi$ is Σ^0_2, whence $V(S_{2,0})$ would turn out to be Σ^0_2, too; contradiction.[30] But what is the complexity of "$v(\sigma_{2,0}) \wedge v(\tau_2)$"? Notationally, it is Σ^0_3; as evaluated in $\mathcal{N}$, it is Δ^0_3; and relative to PA (see footnote 23), it is Δ^0_3, too. For (as a closer look at the formalized completeness theorem would show) the formula "$v(\sigma_{2,0})$" is Σ^0_3, but $T_5 + \mathrm{Con}_{\sigma_{2,0}}$-provably equivalent to a Π^0_3-formula and equivalent to "$\mathrm{Pr}_{v(\sigma_{2,0})}$". Now, since that theory is a subtheory of PA, we have

$$\begin{aligned} \mathrm{PA} \vdash (v(\sigma_{2,0}) \wedge v(\tau_2))(x) &\longleftrightarrow \mathrm{Pr}_{v(\sigma_{2,0})}(x) \wedge \mathrm{Pr}_{v(\tau_2)}(x) \\ &\longleftrightarrow v(\sigma_{2,0})(x) \wedge \mathrm{Pr}_{v(\tau_2)}(x) \end{aligned}$$

—and by what has just been noted, "$v(\sigma_{2,0})(x) \wedge \mathrm{Pr}_{v(\tau_2)}(x)$" is Δ^0_3 in PA. Thus, "$v(\sigma_{2,0}) \wedge v(\tau_2)$" may be taken to be a natural representation of an *axiom-set* of $V(S_{2,0}) \cap V(T_2)$.

After this discussion of theories which were presented as examples for natural theory-internal consistency proofs, let's consider, as sort of a different test case, Feferman's well known representation pa* (see [1]) of an *axiom-set* of PA more thoroughly. pa* is defined thus:

$$\mathrm{pa}^*(x) :\longleftrightarrow \mathrm{pa}(x) \wedge \forall z \leq x \ \mathrm{Con}_{\mathrm{pa}\lceil z} .$$

As far as I know, there is agreement that "$\mathrm{Pr}_{\mathrm{pa}^*}$" is not a natural representation of the theory PA and that "pa*" is not a natural representation of Ax(PA). But it is perhaps not as clear whether "pa*" may not be regarded as a natural representation of the set $\{\psi \mid \mathcal{N} \models \mathrm{pa}^*(\overline{\ulcorner \psi \urcorner})\}$.

Relative to the complexity analysis, "pa*" and "$\mathrm{Pr}_{\mathrm{pa}^*}$" are no natural representations of the sets $\{\psi \mid \mathcal{N} \models \mathrm{pa}^*(\overline{\ulcorner \psi \urcorner})\}$ and PA: for PA is Σ^0_1, whence "pa*" and "$\mathrm{Pr}_{\mathrm{pa}^*}$" needed to be Σ^0_1, too—but they are (Π^0_1 and) Σ^0_2. Given the inductive approach, we have the same result: pa* is certainly not put into the set of natural representations of an *axiom-set* of PA by a base-clause; moreover, there is also no induction step which is applicable to "pa*" in order to make it a natural representation of an *axiom-set* of PA (for "$\mathrm{Pr}_{\mathrm{pa}^*}$", the reasoning is similar).

Yet, what if "pa*" *is* accepted as a natural representation of $\{\psi \mid \mathcal{N} \models \mathrm{pa}^*(\overline{\ulcorner \psi \urcorner})\}$? Besides "refuting" the complexity analysis, this would present a problem for clause (I4) of the inductive approach. That is, the inductive analysis would turn out to be too coarse in the step from an *axiom-set* of

[30] See [9] for a further example where this type of reasoning, employing that the theories intersected are inconsistent with each other, is applied.

a theory to the theory itself. But note now that—since PA *is* consistent—$\{\psi \mid \mathcal{N} \models \mathrm{pa}^*(\overline{\ulcorner \psi \urcorner})\}$ simply is the same set as $\mathrm{Ax}(\mathrm{PA})$. Thus we would have a violation of the following scheme:

> *If α is a natural representation of A and $A = B$,*
> *then α is a natural representation of B.*

I admit that the rejection of that scheme has to be taken into consideration seriously. Yet, without further arguments, I am reluctant to accept the *intensionality* of "α is a natural representation of … ". Therefore, at this point, I am satisfied with the result that both "pa*" and "$\mathrm{Pr}_{\mathrm{pa}^*}$" are no natural representations of the sets of sentences represented by these formulas.

Nevertheless, the fact that "pa*" is no natural representation of $\{\psi \mid \mathcal{N} \models \mathrm{pa}^*(\overline{\ulcorner \psi \urcorner})\}$ in the sense of the inductive approach points to a weakness of the latter: "α is a natural representation of A" it is not explained for sentences α. And a clause of the form

> *If $\alpha(x)$ is a natural representation of A and $\mathcal{N} \models \psi$,*
> *then "$\alpha(x) \wedge \psi$" is a natural representation of $\{\psi \mid \mathcal{N} \models \alpha(\overline{\ulcorner \psi \urcorner}) \wedge \psi\}$*

is certainly inacceptable, as long as "natural representation" is not construed intensionally.[31]

§4. Numerations and closure under necessitation. So far, it seems that the inductive and the complexity approach support each other quite well: *ideally*—with some exceptions noted—the first one generates enough, but probably too many representations as being natural—whence it is the task of the second analysis to eliminate the inacceptable ones. Continuing along this line of thought, I would propose to explain "α is a natural representation of A" by conjoining its three *explicanta* suggested above. Under this assumption, let me consider the question

> Are there examples of theories proving their own consistency naturally?

anew. To answer it, I will investigate more deeply into the numeration approach. In fact, it can be shown for several types of theories T that, if α is a representation of an *axiom-set* of T, it does not numerate that *axiom-set* in T.

To start with, let me present a lemma which is closely related to the main theorem from [9].[32] For its formulation note that, if α numerates A in $\overline{A}$, then

[31] In principle, one could also try to answer these problems by distinguishing between "natural" and "unnatural" *axiom-sets* of theories. In fact, in section 2 I have already called several *axiom-sets* "natural" and, moreover, I think that there are theories, like PA, which are so often presented with just one axiom-system that one may decide to *individuate* them by employing these axiom-systems. But, in general, I have no idea of how to distinguish natural from unnatural *axiom-sets* in a formal way, and I am sceptical if there is any convincing informal motivation for this distinction.

[32] This is: Let $T_1, \ldots, T_k$ be arithmetically definable extensions of PA, having *axiom-sets* with representations $\tau_1, \ldots, \tau_k$ of complexity $\Sigma^0_{i_1}, \ldots, \Sigma^0_{i_k}$ (w.l.o.g. $i_1 \leq i_2 \leq \cdots \leq i_k$), such that $\mathrm{PA} \vdash \forall x(\mathrm{Pr}_q(x) \rightarrow \mathrm{Pr}_{\tau_i}(x))$ $(1 \leq i \leq k)$ and $\forall \varphi$ (φ is a $\Sigma^0_{i_j}$-sentence $\Longrightarrow \mathrm{PA} \vdash \varphi \rightarrow \mathrm{Pr}_{\tau_j}(\overline{\ulcorner \varphi \urcorner})$); let γ

A is closed under Pr_α-necessitation: i.e.,

$$\forall \psi \in \mathrm{L}[A]\ (A \vdash \psi \Longrightarrow A \vdash \mathrm{Pr}_\alpha(\overline{\ulcorner \psi \urcorner})).$$

Lemma 5. If T is a sound extension of S such that $T <_{Co} S$, and if σ is a representation of an *axiom-set* of S, then S is not closed under necessitation with respect to Pr_σ.

Proof. Assume that S is closed under necessitation with respect to Pr_σ, and let ψ in $\mathrm{L}[S]$ be arbitrary. Then $S \vdash \psi \Longrightarrow S \vdash \mathrm{Pr}_\sigma(\overline{\ulcorner \psi \urcorner}) \Longrightarrow T \vdash \mathrm{Pr}_\sigma(\overline{\ulcorner \psi \urcorner})$. Since T is assumed to be sound, $\mathcal{N} \models \mathrm{Pr}_\sigma(\overline{\ulcorner \psi \urcorner})$ follows, whence $S \vdash \psi$.

Now, the function f mapping (the Gödel number of) ψ to (the Gödel number of) "$\mathrm{Pr}_\sigma(\overline{\ulcorner \psi \urcorner})$" is recursive and satisfies

$$S \vdash \psi \iff T \vdash f(\psi).$$

for all $\psi \in \mathrm{L}[S]$. That is, S is many-one reducible to T. But then, $S \leq_{Co} T$ must hold; contradiction.

In [9], I gave a theorem to the effect that completions of PA *never* numerate themselves with formulas also representing *axiom-sets* of them. Here, I present two approximations to a generalization of this result to the case of intersections of theories such that at least one of them is complete (the second is stated (without proof) as Theorem C in [9]).[33]

Theorem 1, Version 1. Assume that $\mathrm{I}\Sigma_1 \subseteq S$, S is complete, T is inconsistent with S and φ is a sentence such that $(S \cap T) + \varphi = S$.[34] If α is a representation of an *axiom-set* A of $S \cap T$ and $S \nvdash \mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner \bot \urcorner})$, then α does not numerate A in $S \cap T$.

Proof. Assume that it does, and let ψ be an arbitrary sentence from $\mathrm{L}[S]$ such that $S \vdash \psi$; then $(S \cap T) + \varphi \vdash \psi$, whence $(S \cap T) \vdash \varphi \rightarrow \psi$, which implies $(S \cap T) \vdash \mathrm{Pr}_\alpha(\overline{\ulcorner \varphi \rightarrow \psi \urcorner})$ by the assumption on α. Therefore, $S \vdash \mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner \psi \urcorner})$.

This shows that S is closed under $\mathrm{Pr}_{\alpha+\varphi}$-necessitation. Now, let by an application of the diagonalization lemma G be the usual *Gödel*-fixed point of "$\mathrm{Pr}_{\alpha+\varphi}$", i.e.,

$$S \vdash G \longleftrightarrow \neg\, \mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner G \urcorner}).$$

If $S \vdash G$, then, being closed under $\mathrm{Pr}_{\alpha+\varphi}$-necessitation, S is inconsistent. Thus, $S \vdash \neg G$, since S is complete. But then, by $\mathrm{Pr}_{\alpha+\varphi}$-necessitation, $S \vdash$

be a sentence from $\mathrm{L}[PA]$ which is consistent with $(T_1 \cap T_2 \cap \cdots \cap T_k)$; and let $(T_1 \cap T_2 \cap \cdots \cap T_k)$ be closed under $\mathrm{Pr}_{\tau_1 \wedge \cdots \wedge \tau_k}$-necessitation, then: $(T_1 \cap T_2 \cap \cdots \cap T_k) + \gamma \nvdash \mathrm{Con}_{\tau_1 \wedge \cdots \wedge \tau_k + \gamma}$.

[33] In case of the second version, it would be interesting to know for which extensions S of PA and for which representations α of *axiom-sets* of S, $S \vdash \mathrm{Rfn}[\alpha\lceil n]$ holds.

[34] Since T inconsistent with S, such sentences exist.

$\mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner\neg G\urcorner})$; and because of the fixed point, $S \vdash \mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner G\urcorner})$. But then $S \vdash \mathrm{Pr}_{\alpha+\varphi}(\overline{\ulcorner\bot\urcorner})$; contradiction.

For the second version, I need an additional concept and two technical lemmata. The new notion is *relative interpretability*, for which I write "$\preceq$"; that is, "$S \preceq T$" is short for "S is relatively interpretable in T". Since it is essentially only one, though central, result of the *theory of relative interpretability* that is employed here, I will omit giving the definition of "$\preceq$" and refer the reader to [1], [4], [6] instead.

Lemma 6. Let $\mathrm{Q} \subseteq S$, α be a numeration of an *axiom-set* A of a theory U in S, and let ψ be a formula; then $(\alpha+\psi)(x)$, i.e., $\alpha(x) \vee x = \overline{\ulcorner\psi\urcorner}$, is a numeration of the *axiom-set* $A \cup \{\psi\}$ of $U + \psi$ in S.

Proof. Because of the assumption on A and α, and since "$\overline{n} = \overline{\ulcorner\psi\urcorner}$" is a Σ^0_0-sentence we have

$$\forall n \in \mathbb{N}\,(n \in A \iff S \vdash \alpha(\overline{n})) \text{ and } \forall n \in \mathbb{N}\,(n = \ulcorner\psi\urcorner \iff S \vdash \overline{n} = \overline{\ulcorner\psi\urcorner}).$$

This implies for all $n \in \mathbb{N}$

$$n \in A \cup \{\ulcorner\psi\urcorner\} \Longrightarrow S \vdash \alpha(\overline{n}) \vee S \vdash \overline{n} = \overline{\ulcorner\psi\urcorner} \Longrightarrow S \vdash \alpha(\overline{n}) \vee \overline{n} = \overline{\ulcorner\psi\urcorner}.$$

On the other hand, assume $(*)$ $S \vdash \alpha(\overline{n}) \vee \overline{n} = \overline{\ulcorner\psi\urcorner}$, but also $n \notin A \cup \{\ulcorner\psi\urcorner\}$. Then $(**)$ $n \notin A$ and $(***)$ $\mathcal{N} \models \overline{n} \neq \overline{\ulcorner\psi\urcorner}$.

$(**)$ implies $S \nvdash \alpha(\overline{n})$, for α is assumed to be a numeration of A in S, and $(***)$ implies $S \vdash \overline{n} \neq \overline{\ulcorner\psi\urcorner}$. But this contradicts $(*)$.

Lemma 7. Assume that PA $\subseteq S$; then for all α, A, φ, U:

if α is a representation of A,
A is an *axiom-set* of U such that $\forall n\ (U \vdash \mathrm{Rfn}[\alpha\lceil n])$, and
φ is a sentence such that $U + \varphi$ is a complete theory,

then α is not a numeration of A in U.

Proof. Applying the diagonalization lemma, let O be the *Orey-sentence* for $\alpha + \varphi$; i.e.,

$$\mathrm{PA} \vdash O \longleftrightarrow \forall x(\mathrm{Con}_{(\alpha+\varphi+O)\lceil x} \rightarrow \mathrm{Con}_{(\alpha+\varphi+\neg O)\lceil x}).^{35}$$

Then, since $\forall n\ (U \vdash \mathrm{Rfn}[\alpha\lceil n])$,

$$\forall n\ U + \varphi + O \vdash \mathrm{Con}_{(\alpha+\varphi+O)\lceil n},$$

Therefore,

$$\forall n\ U + \varphi + O \vdash \mathrm{Con}_{(\alpha+\varphi+\neg O)\lceil n}\,. \tag{*}$$

[35] "$\sigma\lceil z(x)$" is defined as "$\sigma(x) \wedge x \leq z$".

Following [1], let $(\alpha + \varphi + \neg O)^*$ be defined by

$$(\alpha + \varphi + \neg O)^*(x) :\longleftrightarrow \big(\alpha(x) \vee x = \overline{\ulcorner \varphi \urcorner} \vee x = \overline{\ulcorner \neg O \urcorner}\big) \wedge \\ \forall z \leq x \;\; \mathrm{Con}_{(\alpha+\varphi+\neg O)\restriction z} \,.$$

Assume that α is a numeration of A in U. Then:

(i) $U + \varphi + O \vdash \mathrm{Con}_{(\alpha+\varphi+\neg O)^*}$,
(ii) for all $\psi \in \mathrm{L}[U]$
$\big((U + \varphi + \neg O) \vdash \psi \Longrightarrow (U + \varphi + O) \vdash \mathrm{Pr}_{(\alpha+\varphi+\neg O)^*}(\overline{\ulcorner \psi \urcorner})\big)$.

For (i), see the proof of Theorem 5.9 in [1] (which does not presuppose that $U + \varphi + \neg O$ is consistent or r.e.).

For (ii), one first shows that $(**)$ for all $\psi \in \mathrm{L}[U]$,

$$\psi \in A \cup \{\varphi, \neg O\} \Longrightarrow (U + \varphi + O) \vdash \mathrm{Pr}_{(\alpha+\varphi+\neg O)^*}(\overline{\ulcorner \psi \urcorner}).$$

Thus, let $\psi \in A \cup \{\varphi, \neg O\}$; then $U \vdash (\alpha + \varphi + \neg O)(\overline{\ulcorner \psi \urcorner})$ follows by Lemma 6. With $(*)$, this yields $(U + \varphi + O) \vdash (\alpha + \varphi + \neg O)^*(\overline{\ulcorner \psi \urcorner})$, and therefore

$$(U + \varphi + O) \vdash \mathrm{Pr}_{(\alpha+\varphi+\neg O)^*}(\overline{\ulcorner \psi \urcorner}).$$

Now, (ii) can be proved from $(**)$ by an induction on the length of proof in $(U + \varphi + \neg O)$.

From (i) and (ii) it follows by the proof of Theorem 6.1 from [1] that

$$(U + \varphi + \neg O) \preceq (U + \varphi + O). \tag{1}$$

Furthermore, the fixed-point yields

$$\mathrm{PA} \vdash \neg O \longrightarrow \forall x (\mathrm{Con}_{(\alpha+\varphi+\neg O)\restriction x} \rightarrow \mathrm{Con}_{(\alpha+\varphi+O)\restriction x}).$$

Just like before, this implies

$$\forall n \;\, U + \varphi + \neg O \vdash \mathrm{Con}_{(\alpha+\varphi+O)\restriction n},$$

and therefore, employing $(\alpha + \varphi + O)^*$, it can be shown that

$$(U + \varphi + O) \preceq (U + \varphi + \neg O). \tag{2}$$

Now $U + \varphi$ is consistent, whence $(U + \varphi + O)$ is consistent or $(U + \varphi + \neg O)$ is consistent. Since relative interpretability implies relative consistency, we have in both cases by (1) and (2) that $(U + \varphi + O)$ and $(U + \varphi + \neg O)$ are consistent; but this contradicts the completeness of $U + \varphi$.

THEOREM 1, VERSION 2. Assume that $\mathrm{PA} \subseteq S$, S is complete and T not a proper subtheory of S; then for all α, A:

(+) if α is a representation of A, A is an *axiom-set* of $S \cap T$ such that $\forall n \; (S \cap T \vdash \mathrm{Rfn}[\alpha\restriction n])$,

then α is not a numeration of A in $S \cap T$.

Proof. (a) Assume $S = T$; then $S \cap T$ is complete. In this case, the claim follows from Theorem 1 in [9].

(b) Assume $S \neq T$; since S is complete, it is inconsistent with T. Thus there is a sentence φ in L[PA] satisfying $S \vdash \varphi$, $T \vdash \neg\varphi$. Let $U := S \cap T$: then, by $(+)$ A is an *axiom-set* of U with representation α such that $\forall n$ $(U \vdash \text{Rfn}[\alpha\lceil n])$, and such that $U + \varphi (= (S \cap T) + \varphi = S)$ is complete. Thus, Lemma 7 yields that α is not a numeration of A in $S \cap T$.

To close, let me summarize this paper's results for the question whether there are counterexamples to analogues of Gödel's second theorem to non r.e. theories extending (theories like) PA. In section 2, I have presented a wide range of theories T which actually do this; moreover, the arithmetized consistency assertions for these T are formulated with representations of *axiom-sets* of T which are intuitively natural.[36] All of these representations were also natural in the light of the inductive analysis, but only some of them from the perspective of the complexity analysis of "natural representation". Now that we have some results concerning the numeration approach, the picture is as follows:

By Theorem 2 of [9] (in principle), the representations given for $S_1 \cap T_1$ and $S_{2,k} \cap T_2$ do not numerate these theories in themselves. If S_3 is r.e., it's representation pa_ω is too complex; the same type of result holds for S_4. If S_3 fails to be r.e., Lemma 5 applies and pa_ω is no numeration of an *axiom-set* of PA_ω in PA_ω. The representation from example 5 violates the complexity condition, and the one from example 7 is excluded as being natural by Theorem 1, Version 1.[37] Finally, $v(\tau_5)$ (from example 6) is no numeration of $V(T_5)$ in $V(T_5)$ by Theorem 1 from [9].—From the examples discussed in this text, S_4 remains as a serious candidate for a theory which proves its own consistency naturally while fulfilling all adequacy conditions presented here.

REFERENCES

[1] S. Feferman, *Arithmetization of metamathematics in a general setting*, ***Fundamenta Mathematicae***, vol. 49 (1960), pp. 35–92.

[2] ———, *Transfinite recursive progressions of axiomatic theories*, ***The Journal of Symbolic Logic***, vol. 27 (1962), pp. 259–316.

[3] A. Grzegorczyk, A. Mostowski, and C. Ryll-Nardzewski, *The classical and the ω-complete arithmetic*, ***The Journal of Symbolic Logic***, vol. 23 (1958), pp. 188–206.

[4] P. Hájek and P. Pudlak, ***Metamathematics of First-Order Arithmetic***, Springer, Berlin, 1993.

[5] R. Kaye, ***Models of Peano Arithmetic***, Clarendon Press, Oxford, 1991.

[36] Probably with the exception of the representation from example 5.

[37] This holds under the assumption that not both $V(S_{2,0}) \vdash \text{Pr}_{(v(\sigma_{2,0}) \wedge v(\tau_2)) + \varphi}(\overline{\ulcorner \bot \urcorner})$ and $V(T_2) \vdash \text{Pr}_{(v(\sigma_{2,0}) \wedge v(\tau_2)) + \neg\varphi}(\overline{\ulcorner \bot \urcorner})$ (φ being a sentence such that $(V(S_{2,0}) \cap V(T_2)) + \varphi = V(S_{2,0})$ and $(V(S_{2,0}) \cap V(T_2)) + \neg\varphi = V(T_2)$). At present, I cannot exclude this possibility.

[6] P. Lindström, ***Aspects of Incompleteness***, Springer, Berlin, 1997.

[7] G. H. Müller, *Über die unendliche Induktion*, ***Infinitistic Methods***, Pergamon Press, Oxford, London, NY, Paris, 1961, pp. 75–95.

[8] K. G. Niebergall, ***Zur Metamathematik nichtaxiomatisierbarer Theorien***, CIS, München, 1996.

[9] ———, *On the limits of Gödel's second incompleteness theorem*, ***Argument und Analyse. Proceedings of GAP4*** (C. U. Moulines and K. G. Niebergall, editors), Mentis, 2002, pp. 109–136.

[10] C. Smorynski, *The incompleteness theorems*, ***Handbook of mathematical logic*** (J. Barwise, editor), North-Holland, 1977, pp. 821–865.

[11] ———, ***Self-reference and Modal Logic***, Springer, Berlin, 1985.

[12] A. Tarski, A. Mostowski, and R. M. Robinson, ***Undecidable Theories***, North-Holland, Amsterdam, 1953.

[13] A. Visser, *The formalization of interpretability*, ***Studia Logica***, vol. 50 (1991), pp. 81–105.

[14] D. Willard, *Self-verifying axiom systems, the incompleteness theorem and related reflection principles*, ***The Journal of Symbolic Logic***, vol. 66 (2001), pp. 536–596.

SEMINAR FÜR PHILOSOPHIE, LOGIK UND WISSENSCHAFTSTHEORIE
PHILOSOPHIE-DEPARTMENT
LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
LUDWIGSTR. 31, D-80539 MÜNCHEN, GERMANY
E-mail: kgn@lrz.uni-muenchen.de