Faith & Falsity

A Study of Faithful Interpretations and False Σ^0_1-Sentences

Albert Visser

Department of Philosophy, Utrecht University

Heidelberglaan 8, 3584 CS Utrecht, The Netherlands

email: Albert.Visser@phil.uu.nl

April 3, 2003

Abstract

A theory T is trustworthy iff, whenever a theory U is interpretable in T, then it is faithfully interpretable. In this paper we provide a characterization of trustworthiness. We provide a simple proof of Friedman's Theorem that finitely axiomatized, sequential, consistent theories are trustworthy. We provide an example of a theory whose schematic predicate logic is complete Π^0_2.

Key words: Rosser arguments, faithful interpretations, sequential theories, Σ-soundness

MSC2000 codes: 03B52, 03F25, 03F30, 03F45, 03H13

Contents

1 Introduction

Let's begin with a definition.

Definition 1.1 A theory T is *trustworthy* if every U interpretable in T is also faithfully interpretable in T. ❑

Thus our trustworthiness is the trustworthiness of someone who is, in principle, able to truly tell a story without false embellishments. Trustworthiness is a peculiar notion that has nothing to do with strength. It has to do with the constraint a theory puts on the available linguistic means. In Section 6 we will probe deeper into the true and proper nature of trustworthiness. This paper is a study of trustworthiness. We aim to show that the notion of trustworthiness is interesting both in its own right and by its connection to other notions.

1.1 Contents of the Paper

Three central results form the core of the paper. The first is a characterization of trustworthiness. This characterization is provided in Section 5.

As the second central result, we will reprove Friedman's Theorem concerning trustworthiness. The theorem is reported in Craig Smoryński's paper [Smo85a] (Theorem 3, on p224). The theorem states that finitely axiomatized, adequate (sequential[1]), consistent theories are trustworthy. The proof of the result is provided in Section 5. Friedmans' Theorem will be proved as a consequence of our characterization and of a theorem that is proved in Section 4. In fact, the results of Section 4 make a modest strengthening of Friedman's result possible.

Our third central result is the description of trustworthiness in terms af an adjunction between the preorder of faithful interpretability and the preorder of interpretability. This result is proved in Section 6.

An important method used in the paper is the use of the FGH Theorem, which approximately says that we can prove the following principle in Elementary Arithmetic. Let T be a theory into which a suitable fragment of Arithmetic can be interpreted. Then, for any Σ^0_1-sentence S, there is a Σ^0_1-sentence R, such that $(S \vee \mathsf{incon}(T))$ is equivalent to $\Box_T R$. I.o.w. if T is consistent then S is equivalent to a T-provability statement. Since the FGH Theorem plays such an important role, I devote Section 3 to an extensive discussion of it and its applications.

A side result with some independent interest is contained in appendix C. We give an example of a theory whose schematic logic is complete Π^0_2.

1.2 Prerequisites

Most of what is needed to understand the paper is contained in the textbook [HP91].

[1] We will use *sequential* instead of *adequate* in this paper.

1.3 History of the Paper

The present paper is a sequel of [Vis93]. In that work a somewhat sharper version of Theorem 4.1 of the present paper was proved. The present proof is, however, considerably simpler. The article [Vis93], was the result of reflecting on Jan Krajíček's [Kra87]. In that paper Krajíček studies Viteslav Švejdar's question "When is it consistent for inconsistency proofs to lie between cuts?". In other words, for which theories T and for which T-cuts I and J is the theory $T + \mathsf{con}^J(T) + \mathsf{incon}^I(T)$ consistent? Krajíček proves that for every finitely axiomatized, sequential and consistent theory T, and for every T-cut I, we can find a T-cut J such that Švejdar's question has a positive answer for T, I, J.

Neither Krajíček nor I noted that Krajíček's Theorem is an immediate consequence of Friedman's Theorem on trustworthiness.[2] I only realized this recently after Harvey Friedman reminded me of his result in e-mail correspondence. It turns out that in the other direction, the methods of [Vis93] yield a proof of Friedman's Theorem. This paper reports this proof.

1.4 Acknowledgements

I thank Lev Beklemishev and Volodya Shavrukov for providing me with pointers to the literature. I thank Lev also for his comments on the penultimate version of the paper. I am grateful to Harvey Friedman who reminded me of his theorem. I thank Warren Goldfarb and Volodya Shavrukov for e-mails clarifying the history of the FGH Theorem.

2 Arithmetization

In this section we introduce some basic notions and conventions.

2.1 Theories and Interpretations

Theories in this paper are theories of first order predicate logic. Unless stated otherwise, we will assume that theories have an axiom set that is p-time decidable. Interpretations between theories are relative interpretations. For a description of the notion of relative interpretation, see the classical [TMR53], or e.g. [Vis98]. We write:

- $\mathcal{K} : T \rhd U$, for: $\mathcal{K}$ is an interpretation of U in T.
- $T \rhd U$, for $\exists \mathcal{K}\ \mathcal{K}{:}T \rhd U$.

We will be interested in theories in which a sufficiently large fragment of arithmetic is relatively interpretable. Let us fix a weak, finitely axiomatized, arithmetical theory F. Our theory has as language, the arithmetical language with $0, S, +, \times, \leq$. The theory F is axiomatized by Robinson's Arithmetic Q plus

[2] See remark 5.7 of the present paper.

axioms that $\leq$, is linear, plus the axiom $x \leq \mathsf{S}y \leftrightarrow (x \leq y \vee x = \mathsf{S}y)$.[3] We use F instead of Q, because it is pleasant to have some important properties of the Rosser ordering in one's simplest theory.

The theory F is interpretable in Q on a definable initial segment I. See [HP91], pp 366–371. We comment on some details in our appendix D.

To *numerize* a theory T is to specify an interpretation $\mathcal{N}$ such that $\mathcal{N} : T \rhd \mathsf{F}$. Thus, a theory T is *numerizable* if $T \rhd \mathsf{F}$. We will also need the notion of *numerized theory*. A numerized theory $\mathcal{T}$ is a pair $\langle T, \mathcal{N}\rangle$, where $\mathcal{N} : T \rhd \mathsf{F}$. The numerized theory $\langle T, \mathcal{N}\rangle$ is a *numerization* of the numerizable theory T. In the context of numerized theories $\mathcal{T}$, the variables $x, y, z, \ldots$ will range over the numbers provided by $\mathcal{N}$. Thus, e.g. $\forall x \ldots$ will mean $\forall x\, (\delta_{\mathcal{N}}(x) \rightarrow \ldots)$. We will use ξ, η, $\ldots$ for general variables. We will use $\mathcal{T} + A$ for $\langle T + A, \mathcal{N}\rangle$, etc.

We will be sloppy between *numerizable* and *numerized* in the case of 'explicitly arithmetical' theories, like PA. Officially, PA is a numerizable theory. However, we will confuse it with the numerized theory $\langle \mathsf{PA}, \mathsf{id}\rangle$, where id is the identity interpretation.

We will fix an arithmetization of metamathematical notions in the language of F. The arithmetization is supposed to be efficient so that we can verify all relevant facts in Buss' S^1_2. See e.g. [Bus86] or [HP91].[4] We will write $\Box_U A$ ($\Box_{\mathcal{U}} A$), for $\mathsf{prov}_U(\underline{\#A})$ ($\mathsf{prov}_{\mathcal{U}}(\underline{\#A})$). The use of $\Box_U$ ($\Box_{\mathcal{U}}$) will be only meaningful inside a numerized theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$. The formalization of an outer $\Box$ will always be in the designated numbers given by $\mathcal{N}$. *So $\Box_U A$ will be a different formula inside $\langle T, \mathcal{N}\rangle$ than inside $\langle T, \mathcal{K}\rangle$, if $\mathcal{N}$ and $\mathcal{K}$ are different.* Boxes inside boxes will take their numerization from the numerized theory corresponding to the first box above in the parse tree. In appendix A this convention is made precise. The convention is best illustrated by some examples.

Example 2.1 Suppose $\mathcal{T} = \langle T, \mathcal{N}\rangle$ and $\mathcal{U} = \langle T, \mathcal{K}\rangle$ are numerized theories.

- '$\mathcal{T} \vdash \forall \xi\, \exists y\, Q(\xi, y)$', where Q is an atomic predicate, means: $T \vdash \forall \xi\, \exists \eta\, (\delta_{\mathcal{N}}(\eta) \wedge Q(\xi, \eta))$.
- '$\mathcal{T} \vdash \Box_{\mathcal{U}} \forall \xi\, \exists y\, R(\xi, y)$', where R is an atomic predicate, means: $T \vdash \Box_U^{\mathcal{N}} \forall \xi\, \exists \eta\, (\delta_{\mathcal{K}}(\eta) \wedge R(\xi, \eta))$.
- '$T \vdash \Box_U A \rightarrow \Box_U B$' is meaningless. There is nothing to tell us from which set of numbers to take the witnesses for $\Box_U$.
- '$T \vdash \Box_{\mathcal{U}} A \rightarrow \Box_{\mathcal{U}} B$' is meaningless. The witnesses for outer $\Box$'s must come from the numerization of T.
- '$\mathcal{T} \vdash \Box_U A \rightarrow \Box_U B$' means: $T \vdash \Box_U^{\mathcal{N}} A \rightarrow \Box_U^{\mathcal{N}} B$.
- '$\mathcal{T} \vdash \Box_U A \rightarrow \Box_U \Box_U A$' is meaningless. Where could the witnesses for the last $\Box_U$ come from?

[3] Our version of Robinson's Arithmetic has $\leq$ as an atomic symbol and includes the axiom $y \leq x \leftrightarrow \exists z\, z + y = x$. See appendix D.

[4] As is well known, we can replace S^1_2 by a variant in the arithmetical language. We assume we are working with this variant.

- '$\mathcal{T} \vdash \Box_{\mathcal{U}} A \to \Box_{\mathcal{U}} \Box_U A$' means: $T \vdash \Box_U^{\mathcal{N}} A \to \Box_U^{\mathcal{N}} \Box_U^{\mathcal{K}} A$.

❑

Schematic letters A, B, range over the expanded language with boxes and two kinds of variables or over the original language. Schematic letters for Σ_1^0-formulas receive the same treatment as boxed formulas: they range of Σ_1^0-formulas relativized to the stipulated numbers.

Free variables in a formula inside a $\Box$ will be treated according to the usual convention so that they are still free in the resulting formula. Thus, $A(x)$ inside a box will really stand for a term that defines the following function: we map the number n to Gödelnumber of the result of substituting the (binary) numeral $\underline{n}$ of n for x in A.[5]

There are various orderings for interpretations of F in a numerizable theory T. The one that is relevant for us is given as follows.

- $E : \mathcal{K} \leq_T \mathcal{N}$ iff E is a T-formula which T-provably gives an initial embedding of the $\mathcal{K}$-numbers into the $\mathcal{N}$-numbers. We omit the subscript if the theory is clear from the context.

We give the clauses for E. To increase readability we use Plus for $+$ and Times for $\times$.

1. $T \vdash \forall \xi \, \forall \eta \, (E(\xi, \eta) \to (\delta_{\mathcal{K}}(\xi) \wedge \delta_{\mathcal{N}}(\eta)))$,
2. $T \vdash \forall \xi \, (\delta_{\mathcal{K}}(\xi) \to \exists \eta \, (\delta_{\mathcal{N}}(\eta) \wedge E(\xi, \eta)))$,
3. $T \vdash \forall \xi \, \forall \eta \, ((E(\xi, \eta) \wedge \eta' \leq_{\mathcal{N}} \eta) \to \exists \xi' \, (E(\xi', \eta') \wedge \xi' \leq_{\mathcal{K}} \xi))$,
4. $T \vdash \forall \xi \, \forall \xi' \, \forall \eta \, \forall \eta' \, ((E(\xi, \eta) \wedge E(\xi', \eta') \wedge \xi =_{\mathcal{K}} \xi') \to \eta =_{\mathcal{N}} \eta')$,
5. $T \vdash \forall \xi \, \forall \xi' \, \forall \eta \, \forall \eta' \, ((E(\xi, \eta) \wedge E(\xi', \eta') \wedge \mathsf{S}_{\mathcal{K}}(\xi, \xi')) \to \mathsf{S}_{\mathcal{N}}(\eta, \eta'))$,
6. $T \vdash \forall \xi \, \forall \xi' \, \forall \xi'' \, \forall \eta \, \forall \eta' \, \forall \eta'' \, ((E(\xi, \eta) \wedge E(\xi', \eta') \wedge E(\xi'', \eta'') \wedge \mathsf{Plus}_{\mathcal{K}}(\xi, \xi', \xi'')) \to \mathsf{Plus}_{\mathcal{N}}(\eta, \eta', \eta''))$,
7. $T \vdash \forall \xi \, \forall \xi' \, \forall \xi'' \, \forall \eta \, \forall \eta' \, \forall \eta'' \, ((E(\xi, \eta) \wedge E(\xi', \eta') \wedge E(\xi'', \eta'') \wedge \mathsf{Times}_{\mathcal{K}}(\xi, \xi', \xi'')) \to \mathsf{Times}_{\mathcal{N}}(\eta, \eta', \eta''))$.

Any provably initial embedding $E : \mathcal{K} \to \mathcal{N}$ can be split into two parts: $E_0 : \mathcal{K} \to I$, and $\mathsf{emb} : I \to \mathcal{N}$. Here E_0 is a provable isomorphism and I is an initial segment of the $\mathcal{N}$-numbers, satisfying F. The embedding emb is the identical embedding of I into $\mathcal{N}$. We will call such an initial segment of $\mathcal{N}$ satisfying F *a T-cut of $\mathcal{N}$*. If we are considering a numerized theory $\mathcal{T} = \langle T, \mathcal{N} \rangle$, then we will call a T-cut of $\mathcal{N}$ *a $\mathcal{T}$-cut*.

A sequential theory is a theory with a good notion of sequence for all objects of the domain of the theory. This notion is due to Pavel Pudlák. See e.g. [Pud85], or [HP91], p151. The notion of *sequential theory* is equivalent to Harvey Friedman's notion of *adequate theory*. (See [Smo85a].) A sequential theory is always numerizable. Here are a few facts about $\leq$ and cuts.

[5] The 'term' mentioned here need not be really a term, but can given as a suitable formula of which the theory proves that it behaves in the desired way.

Fact 2.2 Consider a numerizable theory T. The variables $\mathcal{K}$, $\mathcal{M}$, $\mathcal{N}$ will range over interpretations of F.

1. For any numerized theory $\mathcal{T} = \langle T, \mathcal{N} \rangle$, there is a $\mathcal{T}$-cut I, such that, for standard k, $\mathcal{T} \vdash \forall x{\in}I\, \exists y\, \mathsf{itexp}(x, \underline{k}) = y$.

   Here, $\mathsf{itexp}(x, 0) := x$ and $\mathsf{itexp}(x, m+1) := 2^{\mathsf{itexp}(x,m)}$.

   This theorem is due to Robert Solovay (in an unpublished manuscript "On Interpretability in Set Theories"). Later a sharper version was proved by Pavel Pudlák in [Pud85]: $\mathsf{S}^1_2 \vdash \forall z\, \exists I\, \Box_{\mathcal{T}} \forall x{\in}I\, \exists y\, \mathsf{itexp}(x, |z|) = y$.

   Here $|n| = \mathsf{entier}(^2\log(n))$.Thus $|n|$ is the binary length of n.

2. For any numerized theory $\mathcal{T} = \langle T, \mathcal{N} \rangle$, there is a $\mathcal{T}$-cut I, such that $I : T \rhd (I\Delta_0 + \Omega_1)$. Since I is a cut, Π^0_1-sentences are downwards preserved from $\mathcal{N}$ to I and Σ^0_1-sentences are upwards preserved from I to $\mathcal{N}$.

   This theorem is due to Alex Wilkie. See [HP91], p366-369. See also our remarks in appendix D.

3. Suppose that T is sequential. Then, for all $\mathcal{M}$, $\mathcal{N}$, there is a $\mathcal{K}$ with $\mathcal{K} \leq \mathcal{M}$ and $\mathcal{K} \leq \mathcal{N}$.

   This theorem is due to Pavel Pudlák ([Pud85]).[6] Note that, by 2., we can always assume that $\mathcal{K} : T \rhd I\Delta_0 + \Omega_1$.

4. Suppose I is a $\mathcal{T}$-cut. Then we have: $\mathsf{S}^1_2 \vdash \forall x\, \Box_{\mathcal{T}}\, x \in I$.

   This theorem is the *obis*-principle. It shows that numbers that are *big outside* are always *small inside*. The result is proved e.g. in [WP87].

□

Remark 2.3 Consider a numerized theory $\mathcal{T} = \langle T, \mathcal{N} \rangle$. Let I, J range over $\mathcal{T}$-cuts. We can assign an invariant to $\mathcal{T}$ as follows:

$$\mathsf{li}(\mathcal{T}) := \{A \mid \exists I\, \forall J{\leq}I\, \mathcal{T} \vdash A^J\}.$$

li stands for 'limes inferior'. It is easily seen that, if $\mathcal{T}$ is consistent, then $\mathsf{li}(\mathcal{T})$ is also consistent. We find that $\mathsf{li}(\mathcal{T})$ extends $I\Delta_0 + B\Sigma^0_1 + \{\mathsf{con}_n(\mathsf{F}) \mid n \in \omega\}$. Here $B\Sigma^0_1$ is the Σ^0_1-collection principle:

$$\vdash \forall x{\leq}a\, \exists y\, S_0(y) \to \exists b\, \forall x{\leq}a\, \exists y{\leq}b\, S_0(y),$$

where $S_0 \in \Delta_0$. The formula con_n stands for consistency w.r.t. n-provability. (See Section 4, for an explanation.)

In case T is sequential, by Fact 2.2(3), $\mathsf{li}(\mathcal{T})$ will be independent of the numerization $\mathcal{T}$ of T. Thus, we may write $\mathsf{li}(T)$, when T is sequential. For sequential theories T and U, we find the following.

[6] Our statement is not precisely Pudlák's, who considers a numerized theory and takes $\mathcal{K}$ to be a cut of the designated numbers. The two statements are easily seen to be equivalent.

1. $\mathsf{li}(T)$ extends $I\Delta_0 + B\Sigma_1^0 + \{\mathsf{con}_n(T) \mid n \in \omega\}$.
2. If $\mathsf{li}(T) \subseteq \mathsf{li}(U)$, then T is locally interpretable in U.
3. If T is finitely axiomatized and consistent, then $\mathsf{li}(T)$ is Σ_1^0-sound. (This follows from Theorem 4.1.)

❑

Open Question 2.4 Remark 2.3 suggests the following questions. What are the possible complexities of the li's? Do we have, for sequential T and U, that if T is interpretable in U, then $\mathsf{li}(T) \subseteq \mathsf{li}(U)$? ❑

2.2 Preliminaries to Rosser Arguments

Suppose $A = \exists x\, A_0(x)$ and $B = \exists x\, B_0(x)$. Here A_0 and B_0 are arbitrary formulas of the language of some numerized theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$. Remember that x and y range over the $\mathcal{N}$-numbers. We write:

- $A \leq B :\leftrightarrow \exists x\, (A_0(x) \wedge \forall y < x \neg B_0(y))$,
- $A < B :\leftrightarrow \exists x\, (A_0(x) \wedge \forall y \leq x \neg B_0(y))$.
- If $C = (A \leq B)$, we write $C^\perp$ for $(B < A)$. If $D = (A < B)$, we write $D^\perp$ for $(B \leq A)$.

Formulas of the form $A \leq B$ and $A < B$ are called *witness comparison formulas*. We present some facts about witness comparison formulas.

Fact 2.5 We have:

1. $\mathcal{T} \vdash A \leq B \to A$.
2. $\mathcal{T} \vdash A < B \to A \leq B$.
3. $\mathcal{T} \vdash A \leq B \to \neg\, (B < A)$.
4. $\mathcal{T} \vdash (A \leq B \wedge B \leq C) \to A \leq C$.
5. $\mathcal{T} \vdash A \leq A \to (A \leq B \vee B < A)$.
6. $\mathcal{T} \vdash (A \wedge \neg\, B) \to A < B$.
7. $\mathcal{T} \vdash ((A \to A \leq A) \wedge B) \to (A \leq B \vee B < A)$.
8. $\mathcal{T} \vdash (A < B \vee B \leq A) \leftrightarrow (A \leq B \vee B < A)$.

❑

Proof

We prove (5). Reason in $\mathcal{T}$. Suppose $A \leq A$. This tells us that $\{x \mid A_0(x)\}$ has a smallest element, say x_0. We have $\forall y{<}x_0\ \neg B_0(y)$ or $\exists y{<}x_0\ B_0(y)$. In the first case, we find $A \leq B$, in the second, $B < A$. ❑

In $I\Delta_0$, we can prove the Δ_0-minimum principle. So, $I\Delta_0 \vdash S \to S \leq S$, for $S \in \exists\Delta_0$. In fact, Δ_0-induction is equivalent to this principle, assuming we allow free parameters in S. Similarly, Buss' theory T^1_2 proves the Σ^b_1-minimum principle.[7] So, $\mathsf{T}^1_2 \vdash S \to S \leq S$, for $S \in \exists\Sigma^b_1$. In fact, Σ^b_1-IND is equivalent to this principle, assuming we allow free parameters in S. (See [Bus86], p61, Theorem 24.) Thus, we can draw the following corollary from Fact 2.5(5,7).

Corollary 2.6 Let S be $\exists\Delta_0$ [$\exists\Sigma^b_1$]. Suppose that $\mathcal{N} : T \rhd I\Delta_0$ [$\mathcal{N} : T \rhd \mathsf{T}^1_2$]. Then, $\mathcal{T} \vdash (S \vee A) \to (S \leq A \vee A < S)$. ❑

Note that it follows, from the conclusion of Corollary 2.6, by substituting S for A, that $\mathcal{T} \vdash S \to S \leq S$, which expresses the Δ_0-minimum principle [Σ^b_1-minimum principle], and hence Δ_0-induction [Σ^b_1-IND]. If, in the Σ^b_1-case, we could prove our corollary using S^1_2, it would follow that $\mathsf{T}^1_2 = \mathsf{S}^1_2$, deciding an open problem. However, we can prove a related fact for S^1_2, which is sufficient for some important applications.

Fact 2.7 Let $\mathcal{T}$ be a numerized theory. Let $\Box := \Box_{\mathcal{T}}$. Suppose that S is $\exists\Sigma^b_1$ and that $A = \exists x\, A_0(x)$. We have $\mathsf{S}^1_2 \vdash S \to \Box(S \leq S)$, and, hence, $\mathsf{S}^1_2 \vdash S \to \Box(S \leq A \vee A < S)$. ❑

Proof

Reason in S^1_2. Suppose S. By Σ^b_1-completeness, we find $\Box S_0(x)$, for some x. By the *obis*-principle, we find $\Box S^I$, for any $\mathcal{T}$-definable cut I. By Fact 2.2(2), we can pick I such that it satisfies $I\Delta_0 + \Omega_1$.[8] It follows that in $\Box(S \leq S)^I$ and, thus, $\Box(S \leq S)$. ❑

The following fact is, modulo some insignificant differences, verified in [VV94].

Fact 2.8 Small Reflection Principle. Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be a sequential numerized theory. Suppose that T is either finitely axiomatized or an extension by finitely many axioms of $I\Delta_0 + \Omega_1$ (relativized to $\mathcal{N}$). Let $\Box := \Box_{\mathcal{T}}$. Let S be $\exists\Sigma^b_1$. Let A be any sentence in the language of T. We have:

$$\mathsf{S}^1_2 \vdash S \to \Box(\Box A \leq S \to A).$$

❑

[7] For a description of T^1_2, see [Bus86] or [HP91].

[8] In fact, we need only a sufficiently large finite fragment of $I\Delta_0 + \Omega_1$ here.

We finish this section by providing a verification of Rosser's Theorem in S^1_2 for theories finitely axiomatized over either S^1_2 or $I\Delta_0 + \Omega_1$. The idea of this argument is due to Viteslav Švejdar. See [Šve83].

Theorem 2.9 Fast Rosser Theorem. *Let $\mathcal{T}$ be a sequential numerized theory. Suppose that T is either finitely axiomatized or an extension by finitely many axioms of $I\Delta_0 + \Omega_1$ (relativized to $\mathcal{N}$). Let $\Box := \Box_{\mathcal{T}}$. Let R be such that $\mathsf{S}^1_2 \vdash R \leftrightarrow \Box\neg R \leq \Box R$. We have: $\mathsf{S}^1_2 \vdash (\Box R \vee \Box\neg R) \rightarrow \Box\bot$.*

Proof

Reason in S^1_2. Suppose (a) $\Box R$. By Fact 2.8, we have (b) $\Box((\Box\neg R \leq \Box R) \rightarrow \neg R)$. By Fact 2.7, we have (c) $\Box((\Box\neg R \leq \Box R) \vee (\Box R < \Box\neg R))$. Combining (b) and (c), we find: (d) $\Box\neg R$. Combining (a) and (d), we get $\Box\bot$. The proof from the assumption $\neg R$ is similar. ❑

Note that it follows, by Buss's results, that there is a p-time transformation of a proof of R to a proof of $\bot$, and, similarly, for proofs of $\neg R$.

Open Question 2.10 The restriction on the theories of Theorem 2.9 is somewhat unsatisfactory. So one might ask whether the theorem also holds for non-sequential theories or for sequential theories that are not either finitely axiomatized or finitely axiomatized as extensions of $I\Delta_0 + \Omega_1$.

It is well known that, if S^1_2 did prove "NP=co-NP", then the usual formalization of Rosser's Theorem would work. Thus, a negative answer to our question would entail: $\mathsf{S}^1_2 \nvdash$ NP=co-NP. ❑

3 A Miraculous Argument

Sometimes, in Mathematics, we meet an argument that is utterly simple, and yet has many surprising consequences. The reasoning leading to the FGH Theorem surely qualifies as an example of such an argument. It is a Rosser type argument and, thus, it inherits the inherent mystery of such arguments. It is a simple Rosser type argument, not much more complicated in terms of number of steps than Rosser's original argument, even simpler in terms of the definition of the fixed point. However, the formalization of the FGH Theorem *seems* to ask for more resources than the formalization of Rosser's, as will be explained below.

3.1 The FGH Theorem

Let us first state the FGH Theorem. Let EA be Elementary Arithmetic, i.e. $I\Delta_0 + \mathsf{exp}$. This theory is called EFA in [Smo85a].

Theorem 3.1 *Consider any numerized theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$. Let $\Box := \Box_{\mathcal{T}}$. Let S be Σ^0_1 and let R be such that $\mathsf{Q} \vdash R \leftrightarrow S \leq \Box R$. We have:*

$$\begin{aligned} \mathsf{EA} \vdash (S \vee \Box\bot) &\leftrightarrow (R \vee \Box\bot) \\ &\leftrightarrow \Box R \end{aligned}$$

or, equivalently, $\mathsf{EA} + \mathsf{con}(T) \vdash (S \leftrightarrow R) \wedge (S \leftrightarrow \Box R)$.

'FGH' stands for Friedman–Goldfarb–Harrington. The history is as follows. Around 1976 or very early 1977, Harrington proved a principle very close to the FGH principle. The main difference was that Harrington's sentence R was Π^0_1 and not Σ^0_1. Harvey Friedman saw Harrington's result and realized that one can also get the result for R in Σ^0_1. He wrote down his result in a manuscript "Proof Theoretic Degrees", dated February 1977. An early paper reporting the result is Smoryński's [Smo81], p366. Smoryński refers to Friedman's unpublished manuscript.

Warren Goldfarb rediscovered the principle independently in November 1980. He communicated the result to George Boolos. Boolos then promulgated it to the logic of provability community. Via this channel I learned of it. So I called it *Goldfarb's Principle*. I guess everyone gets due credit in my new name for it: *The FGH Theorem*. Here is the proof.

Proof

Reason in EA.

Step 1. Suppose $S \vee \Box\bot$. We want to derive $R \vee \Box\bot$. If we have $\Box\bot$, we are done. Suppose S. It follows that $R \vee R^\bot$. In the first case, we are again done. In case we have $R^\bot$, we find (a) $\Box R$, since $R^\bot = (\Box R < S)$. Moreover, by Σ^0_1-completeness, we have (b) $\Box R^\bot$. Combining (a) and (b), we obtain $\Box\bot$.

Step 2. Suppose $R \vee \Box\bot$. By Σ^0_1-completeness, we find $\Box R \vee \Box\bot$, hence, $\Box R$.

Step 3. Suppose $\Box R$. We want to derive $R \vee \Box\bot$. We find: $R \vee R^\bot$. Now we may proceed as in step 1.

Step 4. Suppose $R \vee \Box\bot$. We may immediately conclude that $S \vee \Box\bot$. ❑

Remark 3.2 We can also prove $\mathsf{EA} \vdash \Box\neg R^\bot \leftrightarrow \Box R$. Right-to-left is trivial. In the other direction, let I be a $\mathcal{T}$-cut satisfying $I\Delta_0 + \Omega_1$. We have:

$$\begin{array}{lll} \mathsf{EA} \vdash \Box\neg R^\bot & \rightarrow & \Box(\neg R^\bot)^I \\ & \rightarrow & \Box(\Box R \rightarrow R)^I \\ & \rightarrow & \Box(\Box^I R \rightarrow R^I) \\ & \rightarrow & \Box(\Box^I R \rightarrow R) \\ & \rightarrow & \Box R \end{array}$$

The last step is an application of Löb's Theorem for $\langle T, I\rangle$. For a discussion of Löb's theorem with shifting interpretations, see [Vis93], section 4. ❑

An immediate generalization of the FGH theorem is due essentially to Franco Montagna.

Theorem 3.3 *Consider any numerized theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$. Let $\Box := \Box_{\mathcal{T}}$. Let $S(x)$ be Σ^0_1 and let R be such that $\mathsf{Q} \vdash R \leftrightarrow S(\underline{\#R}) \leq \Box R$. We have:*

$$\begin{array}{rcl} \mathsf{EA} \vdash (S(\underline{\#R}) \vee \Box\bot) & \leftrightarrow & (R \vee \Box\bot) \\ & \leftrightarrow & \Box R \end{array}$$

or, equivalently, $\mathsf{EA} + \mathsf{con}(T) \vdash (S(\underline{\#R}) \leftrightarrow R) \wedge (S(\underline{\#R}) \leftrightarrow \Box R)$.

It is easy to see that Rosser's Theorem is an immediate consequence of Montagna's Theorem. We end this subsection, by proving a variant of a part of the FGH Theorem that will be used in Section 4.

Theorem 3.4 *Consider any numerized theory $\mathcal{T}$. Let $\Box := \Box_{\mathcal{T}}$. Let A be $\exists\forall\Delta_0$ and let R be such that $\mathsf{Q} \vdash R \leftrightarrow A \leq \Box R$. Let $B\Sigma^0_1$ be the Σ^0_1-collection principle: $\vdash \forall x{\leq}a\, \exists y\, S_0(y) \rightarrow \exists b\, \forall x{\leq}a\, \exists y{\leq}b\, S_0(y)$, where $S_0 \in \Delta_0$. We have:*

$$\mathsf{EA} + B\Sigma_1 \vdash \Box R \rightarrow (A \vee \Box\bot),$$

or, equivalently, $\mathsf{EA} + B\Sigma_1 + \mathsf{con}(T) \vdash \Box R \rightarrow A$.

Proof

Reason in $\mathsf{EA} + B\Sigma_1$. Suppose $\Box R$. We have $A \leq \Box R$ or $\Box R < A$. In the first case, we may conclude A, and we are done. Suppose $\Box R < A$. This has the form $\exists p\ (\mathsf{proof}(p, \underline{\#R}) \wedge \forall y{\leq}p\ \exists z\ \neg A_0(y,z))$, where A_0 is in Δ_0. By Σ^0_1-collection, our formula is equivalent to:

$$C := \exists p\, \exists x\ (\mathsf{proof}(p, \underline{\#R}) \wedge \forall y{\leq}p\ \exists z \leq x\ \neg A_0(y,z)).$$

Thus, we find: $\Box C$, and, hence, $\Box(R < A)$. I.o.w., $\Box R^{\bot}$. Combining this with our assumption $\Box R$, we find $\Box\bot$ and we are done. ❑

3.2 The FGH Theorem and S^1_2

It is an open problem whether the FGH Theorem can be formalized in S^1_2, even for $S \in \exists\Sigma^b_1$. However for a restricted range of theories we can prove a salient consequence of FGH Theorem.

Theorem 3.5 *Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be a sequential numerized theory. Suppose that T is either finitely axiomatized or an extension by finitely many axioms of $I\Delta_0 + \Omega_1$ (relativized to $\mathcal{N}$). We write $\Box := \Box_{\mathcal{T}}$. Let A be any T-sentence. Let R be such that $\mathsf{Q} \vdash R \leftrightarrow \Box A \leq \Box R$. We have: $\mathsf{S}^1_2 \vdash \Box A \leftrightarrow \Box R$.*

Proof

Reason in S^1_2.

Suppose $\Box A$. By the small reflection principle 2.8, we have (a):

$$\Box((\Box R < \Box A) \to R).$$

By $\Box A$ and Fact 2.7, we have (b) $\Box(\Box A \leq \Box R \vee \Box R < \Box A)$. Combining (a) and (b), we find $\Box R$.

Conversely, suppose $\Box R$. By the small reflection principle 2.8, we have:

$$\Box((\Box A \leq \Box R) \to A),$$

i.e. $\Box(R \to A)$. Ergo, $\Box A$. ❑

3.3 Some Consequences of the FGH Theorem

The proof of Theorem 4.1 is our central application of the FGH Theorem in this paper. We also use it in the proof of Theorem C.7. In this subsection we spell out some more immediate consequences of Theorem 3.1. These consequences are not strictly needed for the rest of the paper. They have, however, heuristic value. Moreover, they are interesting in their own right. For some further information, the reader is referred to [Smo85b], chapter 7.

3.3.1 1-Reducibility

We give a quick proof of a well-known fact.

Theorem 3.6 *Suppose T can be extended to a consistent numerizable theory W. Then, any RE set is 1-reducible to T. A fortiori, T is of Turing degree $\mathbf{0}'$.*

Proof

Clearly, we may assume that W is a finite extension of T, say $W = T + A$. Let $\mathcal{W} = \langle W, \mathcal{N}\rangle$ be a numerization of W. Consider any RE set X with index e. Let R_n be the FGH sentence for the theory $\mathcal{W}$ corresponding to the sentence $S_n := (\{\underline{e}\}\underline{n} \simeq 0)$. Clearly, the mapping $n \mapsto (A \to R_n^{\mathcal{N}})$ is recursive. By the FGH Theorem, formulated externally, we have: $n \in X \Leftrightarrow T \vdash A \to R_n^{\mathcal{N}}$. ❑

3.3.2 Closure under Disjunction

We show that provabilities are closed under disjunction.

Theorem 3.7 *Let $\mathcal{T}$ be a numerized theory. Let $\Box := \Box_{\mathcal{T}}$. For any sentences A and B of the language of T, there is a Σ_1-sentence C such that $\mathsf{EA} \vdash \Box C \leftrightarrow (\Box A \vee \Box B)$.*

Proof

Take $S := (\Box A \vee \Box B)$ in Theorem 3.1. ❑

Note that C can in fact be taken to be $\exists\Pi^b_1$.

3.3.3 Degrees of Provably Deductive Consequence

Let $\mathcal{T}$ be numerized. Let A and B be be sentences of the language of $\mathcal{T}$. Let $\Box := \Box_{\mathcal{T}}$. We define:

- $A \preceq_{\mathcal{T}} B :\Leftrightarrow \mathcal{T} \vdash \Box A \to \Box B$.
- $A \equiv_{\mathcal{T}} B :\Leftrightarrow A \preceq_{\mathcal{T}} B$ and $B \preceq_{\mathcal{T}} A$.

We call $\preceq_{\mathcal{T}}$ *provably deductive consequence* and we call $\equiv_{\mathcal{T}}$ *provably deductive equivalence*. Clearly, these notions yield a degree structure on the sentences of $\mathcal{T}$.

Theorem 3.8 *Each degree of provably deductive equivalence of $\mathcal{T}$ contains a $\exists\Pi^b_1$-sentence.*

Proof

Let γ be such a degree. Suppose $C \in \gamma$. Take $S := \Box C$ in Theorem 3.1. ❑

3.3.4 Smoryński's Theorem

The following application is due to Smoryński. See [Smo81], p366 or [Smo85b], p312.

Theorem 3.9 *Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be numerized theory. Suppose $\mathcal{T} \vdash \mathsf{EA}$. Then, we have, verifiably in EA, that $\mathcal{T}$ is Σ^0_1-sound iff $\mathcal{T}$ is consistent and $\mathcal{T} + \mathsf{con}(T)$ is Σ^0_1-conservative over $\mathcal{T}$.*

Proof

We write $\Box := \Box_{\mathcal{T}}$. Reason in EA.

Suppose $\mathcal{T}$ is Σ^0_1-sound. Let S be in Σ^0_1. Suppose $\Box(\mathsf{con}(T) \to S)$. Then, we find $\Box(S \vee \Box\bot)$. By Σ^0_1-soundness, it follows that $(S \vee \Box\bot)$. Hence, by Σ^0_1-completeness, $\Box S$.

Suppose that $\mathcal{T}$ is consistent and $\mathcal{T} + \mathsf{con}(T)$ is Σ^0_1-conservative over $\mathcal{T}$. Suppose $\Box S$. Applying the first equivalence of the FGH Theorem *inside the* $\Box$, we obtain $\Box(R \vee \Box\bot)$. Ergo, $\Box(\mathsf{con}(T) \to R)$. By Σ^0_1-conservativity, it follows that $\Box R$. We may conclude, now applying the FGH Theorem outside the $\Box$, that S. ❑

Note that the assumption that $\mathcal{T} \vdash \mathsf{EA}$, was only used in the second part of the proof in the 'internal' application of the FGH Theorem. We can extend the result to theories $\mathcal{T}$ such that every $\mathcal{T}$-cut I has a subcut J with $J : T \rhd T$. Examples of such theories are S^1_2, $I\Delta_0 + \Omega_{14} + \mathsf{con}(\mathsf{F})$, $I\Delta_0 + \{\Omega_{n+1} \mid n \in \omega\}$ and $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$,

Theorem 3.10 *Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be numerized theory. Suppose that every $\mathcal{T}$-cut I has a subcut J with $J : T \rhd T$. Then, we have, verifiably in EA, that $\mathcal{T}$ is Σ^0_1-sound iff $\mathcal{T}$ is consistent and $\mathcal{T} + \mathsf{con}(T)$ is Σ^0_1-conservative over $\mathcal{T}$.*

Proof

We replace the second part of the previous proof by the following variation. Suppose that $\mathcal{T}$ is consistent and that $\mathcal{T} + \mathsf{con}(T)$ is Σ^0_1-conservative over $\mathcal{T}$. Suppose $\Box S$. Using Fact 2.2(2),(1), we can find a $\mathcal{T}$-cut J such that (a) $J : T \rhd (I\Delta_0 + T)$ and (b) $\Box(\forall x{\in}J\, \exists y\ 2^x = y)$. By (a), we find $\Box S^J$. Ergo $\Box(R \vee R^\perp)^J$, Hence, $\Box(R \vee (R^\perp)^J)$. Also (c) $\Box((R^\perp)^J \to \Box R)$. Since, in the proof of Σ^0_1-completeness for $\mathcal{T}$, the transformation of the witness x of a Σ^0_1-sentence S' to a proof p of S' is of order 2^{x^m}, for standard m, we get by (b): $\Box((R^\perp)^J \to \Box R^\perp)$. Ergo (d) $\Box((R^\perp)^J \to \Box\perp)$. We may conclude from (c) and (d): $\Box(R \vee \Box\perp)$.

Hence, $\Box(\mathsf{con}(T) \to R)$. By Σ^0_1-conservativity, it follows that $\Box R$. We may conclude, by the FGH Theorem, that S. ❑

Here is a corollary from Theorem 3.9. A theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$ is *reflexive* if it proves for every n the statement $\mathsf{con}(\mathcal{T}_n)$. Here $\mathcal{T}_n$ is the theory axiomatized by $\mathsf{EA}^{\mathcal{N}}$ plus the T-axioms with Gödelnumber less than or equal to n.

Corollary 3.11 Suppose $\mathcal{T}$ is a consistent, numerized, reflexive theory such that $\mathcal{T} \vdash \mathsf{EA}$. Suppose there is an n, such that, for all Σ^0_1-sentences S, whenever $\mathcal{T} \vdash S$, we have $\mathcal{T}_n \vdash S$. Then $\mathcal{T}$ is Σ^0_1-sound. ❑

Proof

Let $\Box := \Box_{\mathcal{T}}$ and $\Box_n := \Box_{\mathcal{T}_n}$. Suppose $\Box_n(\mathsf{con}(\mathcal{T}_n) \to S)$, then, by reflexivity, $\Box S$. Hence, $\Box_n S$. Applying Theorem 3.9 to $\mathcal{T}_n$, we find S. So $\mathcal{T}_n$ is Σ^0_1-sound. Hence, $\mathcal{T}$ is also Σ^0_1-sound. ❑

The above theorem tells us that, if a theory that is consistent, numerized, reflexive and verifies EA, proves a false Σ^0_1-sentence, then it is forced to tell more and more complex lies, i.e., it will prove false Σ^0_1-sentences the proofs of which need more and more axioms.

4 Σ^0_1-Soundness in Potentia

In this section, we prove a theorem that will be the main lemma to our proof that consistent, finitely axiomatized, sequential theories are trustworthy. Let EA^+ be

$I\Delta_0 + \mathsf{supexp}$, where supexp is the axiom stating that the superexponentiation function is total.

Theorem 4.1 *Let $\mathcal{T} := \langle T, \mathcal{N}\rangle$ be a finitely axiomatized, sequential theory. We write $\Box := \Box_{\mathcal{T}}$. There is a $\mathcal{T}$-cut I such that, for all Σ^0_1-sentences S, $\mathsf{EA}^+ \vdash (S \vee \Box\bot) \leftrightarrow \Box S^I$, or equivalently, $\mathsf{EA}^+ + \mathsf{con}(T) \vdash S \leftrightarrow \Box S^I$.*

Before proving our theorem we formulate and prove an immediate corollary.

Corollary 4.2 Let $\mathcal{T} := \langle T, \mathcal{N}\rangle$ be a finitely axiomatized, sequential theory. There is a $\mathcal{T}$-cut I such that $\langle T, I\rangle$ is $(\mathsf{EA}^+ + \mathsf{con}(T))$-verifiably Σ^0_1-sound. ❑

Proof

Let I be the cut promised in theorem 4.1. We have, for any Σ^0_1-sentence S, $\mathsf{EA}^+ + \mathsf{con}(T) \vdash \Box_{\mathcal{T}} S^I \to S$, and hence, $\mathsf{EA}^+ + \mathsf{con}(T) \vdash \Box_{\langle T, I\rangle} S \to S$. ❑

To get the proof of theorem 4.1 going, we need a few preparatory steps. We will apply the FGH Theorem to a restricted proof predicate, where the formulas in the proof are restricted to formulas of a certain complexity. We take as measure of complexity ρ, where $\rho(A)$ is the depth of quantifier changes. This measure is discussed in some detail in [Vis93]. We take Γ_n to be the set of formulas of complexity at most n and Γ^{cl}_n the set of sentences of Γ_n. m-provability will be provability from axioms with Gödelnumber below m, where the formulas occurring in the proof are all in Γ_m.

The notation $A(\underline{k})$ is somewhat misleading. In general we are working in some *interpretation* of number theory. So the term $\underline{k}$ occurs in unwinded relational form. Our measure ρ is designed to be insensitive to such fine points.

Lemma 4.3 $\rho(A(\underline{k}))$ is independent of k. ❑

Proof

Suppose, for simplicity, that we are working with tally-numerals. $A(\underline{k})$ in $\mathcal{T}$ could look like this:

$$\exists x_0 \ldots \exists x_k \, (0^{\mathcal{N}}(x_0) \wedge S^{\mathcal{N}}(x_0, x_1) \wedge \ldots \wedge S^{\mathcal{N}}(x_{k-1}, x_k) \wedge A(x_k)).$$

The complexity of this formula is $\max(\rho(0^{\mathcal{N}}(x)), \rho(S^{\mathcal{N}}(x, y)), \rho(A(x))) + 1$. This formula is clearly estimated by $\rho(A(x)) + c$, for a fixed standard c. Similar reasoning works for efficient numerals based e.g. on binary notations. ❑

Here is a fundamental lemma about $\Box_n$.

Lemma 4.4 Suppose that $\mathcal{T} := \langle T, \mathcal{N}\rangle$ is a finitely axiomatized theory. Let $\Box$ and $\Box_m$ be the provability and the m-provability predicates of $\mathcal{T}$. We have, for any T-sentence A and $k > \rho(A)$ and k larger than the complexities of the axioms of T, $\mathsf{EA}^+ \vdash \Box_k A \leftrightarrow \Box A$. ❑

Proof

The left-to-right direction is obvious. To prove the right-to-left direction, reason in EA^+. Suppose $\Box A$. We can, using supexp, find a cutfree proof in predicate logic of $C \to A$, where C is the conjunction of the T-axioms. See [HP91], part V, chapter 5, for details. By the subformula property, this proof is also an k-proof. ❑

Note that we used the fact that T is finitely axiomatized in an essential way in the proof.

Open Question 4.5 Is it possible to replace, in the usual superexponential estimate of the growth involved in cut elimination, the usual measure of complexity (depth of connectives) by ρ, i.e. depth of quantifier changes? ❑

Lemma 4.6 Let $\mathcal{T} := \langle T, \mathcal{N} \rangle$ is a finitely axiomatized. Let $\Box$ and $\Box_m$ be the provability and the m-provability predicates of $\mathcal{T}$. Consider a Σ^0_1-sentence S. We can find R_m such that $\mathsf{Q} \vdash R_m \leftrightarrow S \leq \Box_m R_m$, by the Gödel Fixed Point Lemma. Note that $\rho(R_m) := \rho(S) + c$, for a standard c which is independent of m. Choose $n > \rho(S) + c$. We have: $\mathsf{EA}^+ \vdash (S \vee \Box\bot) \leftrightarrow \Box R_n$. ❑

Proof

We want to apply the FGH Theorem. To do this we must verify that the steps in the proof go through for our n-provability. Note e.g. that n is large enough to have: $\mathsf{EA} \vdash R_n \to \Box_n R_n$ and $\mathsf{EA} \vdash R_n^\bot \to \Box R_n^\bot$. Thus, we have: $\mathsf{EA} \vdash (S \vee \Box_n\bot) \leftrightarrow \Box_n R_n$. Now apply Lemma 4.4. ❑

Our proof strategy will be to provide a cut I, such that, EA^+-verifiably, we have $\Box R_n \leftrightarrow \Box S^I$. Then we may apply Lemma 4.6. To get the desired result, we need a reflection principle.

Lemma 4.7 Let $\mathcal{U} := \langle U, \mathcal{M} \rangle$ be any sequential theory. Let $\Box$ be $\mathcal{U}$-provability and let $\Box_n$ be $\mathcal{U}$-n-provability. For any n, we can find an $\mathcal{U}$-cut J such that $\mathsf{EA} \vdash \forall A \in \Gamma^{\mathsf{cl}}_n \ \ \Box(\Box^J_n A \to A)$. ❑

Proof

This is Fact 2.4.5(ii) of [Vis93]. The idea is that, in $\mathcal{U}$, we can define a satisfaction predicate for Γ_n and prove Γ_n-reflection by replacing induction over proof length by the use of a definable cut. ❑

The next lemma is nearly the theorem we are aiming to prove. The only defect is that I is still dependent on $\rho(S)$.

Lemma 4.8 Let $\mathcal{T} := \langle T, \mathcal{N} \rangle$ be a finitely axiomatized, sequential theory. We write $\Box := \Box_{\mathcal{T}}$. For any Σ^0_1-sentence S, there is a $\mathcal{T}$-cut I such that, $\mathsf{EA}^+ \vdash (S \vee \Box\bot) \leftrightarrow \Box S^I$, or equivalently, $\mathsf{EA}^+ + \mathsf{con}(T) \vdash S \leftrightarrow \Box S^I$. The cut I depends only on $\rho(S)$. ❑

Proof

Take n and R_n as in Lemma 4.6. Let $R := R_n$. We have, by Lemma 4.6, (a) $\mathsf{EA}^+ \vdash (S \vee \Box\bot) \leftrightarrow \Box R$. Choose a reflecting $\mathcal{T}$-cut I for $\Box_n$ as in Lemma 4.7. By Fact 2.2(2), we can choose I in such a way that it verifies Δ_0-induction. Note that I will only depend on $\rho(S)$.

The left-to-right direction is immediate by the *obis*-principle. We treat the other direction. By (a), it is sufficient to show that $\mathsf{EA}^+ \vdash \Box S^I \rightarrow \Box R$.

Reason in EA^+. Suppose $\Box S^I$. Since we have Δ_0-induction in I, it follows that $\Box(S \leq \Box_n R \vee \Box_n R < S)^I$ and so $\Box((S \leq \Box_n R)^I \vee (\Box_n R < S)^I)$. The first disjunct is equivalent to R^I, which implies R. To the second disjunct we apply the reflection principle from Lemma 4.7 to infer R. Thus, we obtain $\Box R$. ❑

We want to make the cut I independent of the Σ^0_1-sentence S. The problem is that Σ^0_1-sentences may have arbitrarily large ρ-complexities. If we would have $\mathcal{N} : T \rhd \mathsf{EA}$, there would be no problem, since we have $\mathsf{EA} \vdash S \leftrightarrow \mathsf{true}_\Sigma(\underline{\#S})$, where true_Σ is the ordinary Σ^0_1-truth predicate, which is itself given by a Σ^0_1-formula. All sentences of the form $\mathsf{true}_\Sigma(\underline{\#S})$ have some complexity below a fixed finite n. We can use the idea even in the absence of EA by making our cut smaller. Here is another lemma.

Lemma 4.9 Let $S = \exists x S_0(x)$, where $S_0 \in \Delta_0$. Let the truth predicate be of the form $\exists y\, \mathsf{true}_{\Sigma,0}(y, z)$, where $(\mathsf{true}_{\Sigma,0}(y, z)) \in \Delta_0$. There is a fixed standard k, such that $\mathsf{S}^1_2 \vdash (S_0(x) \wedge 2^{x^{\underline{k}}}\!\downarrow) \rightarrow \exists y{\leq}2^{x^{\underline{k}}}\, \mathsf{true}_{\Sigma,0}(y, \underline{\#S})$. ❑

Proof

The proof is by inspecting the usual EA-proof of $S \rightarrow \mathsf{true}_\Sigma(\underline{\#S})$. See e.g. [HP91], part C, chapter 5(b), for a detailed presentation. ❑

Here is the proof of Theorem 4.1.

Proof

Let J be the cut provided by Lemma 4.8 for the complexity of the Σ-truth-predicate. Let I a shorter cut, such that $\mathcal{T} \vdash \forall x{\in}I\ \ 2^x \in J$.

Let a Σ^0_1-sentence S be given. The left-to-right direction is immediate, using the *obis*-principle. We treat the direction from right-to-left. Take $S_* := \mathsf{true}_\Sigma(\underline{S})$. By Lemma 4.8, we get $\mathsf{EA}^+ \vdash (S_* \vee \Box\bot) \leftrightarrow \Box S^J_*$. By Lemma 4.9, we find $\mathcal{T} \vdash S^I \rightarrow S^J_*$. Thus, we have:

$$\begin{array}{rlll} \mathsf{EA}^+ \vdash & \Box S^I & \rightarrow & \Box S^J_* \\ & & \rightarrow & S_* \vee \Box\bot \\ & & \rightarrow & S \vee \Box\bot \end{array}$$

So we are done. ❑

Open Question 4.10 Can one find a numerized, non-sequential, finitely axiomatized theory for which there is a false Σ^0_1-sentence which is provable on every definable cut? ❑

We draw an obvious corollary.

Corollary 4.11 Suppose $\mathcal{T}$ is consistent, finitely axiomatized and sequential. Then there are a $\mathcal{T}$-cut I and a model $\mathcal{M}$ of $\mathcal{T}$ such that, in $\mathcal{M}$, witnesses of Σ^0_1-sentences are either in the initial segment of the $\mathcal{T}$-numbers isomorphic to ω or not in I. ❑

Proof

Choose I as in Theorem 4.1. Clearly, $U := T + \{\neg S^I \mid \mathbb{N} \not\models S\}$ is consistent. Take $\mathcal{M}$ a model of U. ❑

Note that Corollary 4.11, in its turn, directly implies Theorem 4.1. Another immediate corollary is as follows. This corollary is about the limes inferior of a sequential theory T. The notion of limes inferior of a sequential theory T or $\mathsf{li}(T)$ was introduced in remark 2.3.

Corollary 4.12 Let T be a consistent, sequential, finitely axiomatized theory. Then $\mathsf{li}(T)$ is Σ^0_1-sound. ❑

We can extend Lemma 4.8 partly to a wider formula class.

Definition 4.13 Consider any numerized theory $\mathcal{T}$. Let $B := \exists x\ B_0(x)$ be a formula of the language of $\mathcal{T}$. Let I be a $\mathcal{T}$-cut. We write $B^{[I]}$ for $\exists x{\in}I\ B_0(x)$ (or: $B < \exists x\, x \notin I$). ❑

Theorem 4.14 *Let $\mathcal{T} := \langle T, \mathcal{N}\rangle$ be a finitely axiomatized, sequential theory. We write $\Box := \Box_{\mathcal{T}}$. For any $\exists\forall\Delta_0$-sentence A, there is a $\mathcal{T}$-cut I such that, $\mathsf{EA}^+ + B\Sigma_1 \vdash \Box A^{[I]} \to (A \vee \Box\bot)$, or equivalently,*

$$\mathsf{EA}^+ + B\Sigma_1 + \mathsf{con}(T) \vdash \Box A^{[I]} \to A.$$

The cut I depends only on $\rho(A)$.

Proof

Take R as in Theorem 3.4, with $\Box_n$, for a suitably large n, substituted for $\Box$. We find, using cut elimination, from Theorem 3.4:

$$\mathsf{EA}^+ + B\Sigma_1 \vdash \Box R \to (A \vee \Box\bot).$$

Let I be an n-reflecting $\mathcal{T}$-cut satisfying $I\Delta_0$. It is sufficient to show in $\mathsf{EA}^+ + B\Sigma_1$ that $\Box A^{[I]}$ implies $\Box R$.

Reason in $\mathsf{EA}^+ + B\Sigma_1$. Suppose $\Box A^{[I]}$. Since

$$\Box(\Box_n R \to \Box_n R \leq \Box_n R)^I,$$

it follows, by Fact 2.5(7), that $\Box(A^{[I]} \leq \Box_n^I R \vee \Box_n^I R < A^{[I]})$.[9] Clearly, the first disjunct is $\mathcal{T}$-equivalent to $R^{[I]}$, and, thus, implies in $\mathcal{T}$ that R. Moreover, the second disjunct implies in $\mathcal{T}$ that $\Box_n^I R$. Hence, since I is n-reflecting, the second disjunct implies R in $\mathcal{T}$. Thus, we find (outside of $\mathcal{T}$): $\Box R$. ❑

We can extend Theorem 4.1 to a larger class of theories.

Theorem 4.15 *Let T be a consistent, sequential, finitely axiomatized theory. Suppose that T and U are mutually interpretable. Then there is a Σ_1^0-sound numerization $\mathcal{U} = \langle U, \mathcal{P}\rangle$ of U.*

Note that U need not be sequential! Before proving the theorem we need a lemma, which is a strengthening of Löb's Theorem.

Lemma 4.16 Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be a numerized, consistent, sequential, finitely axiomatized theory. . Let I be a $\mathcal{T}$-cut and let A be a sentence of the language of T. Then there is a k such that

$$I\Delta_0 + \Omega_1 \vdash \Box_{\mathcal{T}}(\Box_{T,k}^I A \to A) \to \Box_{\mathcal{T}} A.$$

The number k depends only on the complexities of the axioms of T, the complexity of $\mathcal{N}$, the complexity of I and the complexity of A. Our complexity measure here is ρ, i.e. depth of quantifier changes. ❑

The lemma is a special case of Theorem 4.2 of [Vis93]. We turn to the proof of Theorem 4.15.

Proof

Suppose $\mathcal{K} : T \rhd U$ and $\mathcal{M} : U \rhd T$. Note that $\mathcal{N}' := \mathcal{N}\mathcal{M}\mathcal{K}$ is an interpretation of F in T. (We write composition in the order of application here.) By Fact 2.2(3), there is a $\mathcal{T}$-cut J that is $\mathcal{T}$-provably isomorphic with a T-cut J' of $\mathcal{N}'$. By Fact 2.2(2), we may assume that J satisfies $I\Delta_0+\Omega_1$. Let K be the ρ-complexity of the Σ_1^0-truth predicate. By the external form of Lemma 4.16, we can find a k such that, for any $A \in \Gamma_{K+n}$, if $\mathcal{T} \vdash \Box_{T,k}^J A \to A$, then $\mathcal{T} \vdash A$. Here n is a sufficiently large number.

By Lemma 4.7, we can find a $\mathcal{T}$-cut I^* such that $\mathcal{T} \vdash \Box_{T,k}^{I^*} B \to B$, for any $B \in \Gamma_k$. Let I be a subcut of I^* such that $\mathcal{T} \vdash \forall x{\in}I\ 2^x \in I^*$. By Fact 2.2(2), we may choose I^* and I such that they satisfy $I\Delta_0 + \Omega_1$. Consider any Σ_1^0-sentence S. Let $S_0 := \mathsf{true}_\Sigma(\underline{\#S})$. We have, by Lemma 4.9, $\mathcal{T} \vdash S^I \to S_0^{I^*}$.

[9] Note that, to apply the verbatim statement of Fact 2.5(7) we have to shift to the theory $\langle T, I\rangle$ first and, then, shift back to $\mathcal{T}$. Alternatively, we can just run through the proof again for the modified statement.

Let R be such that $\mathsf{F} \vdash R \leftrightarrow S_0 \leq \Box_{\mathcal{T},k} R$. We have:

$$\begin{array}{rrcl} \mathcal{T} \vdash & S^I & \rightarrow & S_0^{I^*} \\ & & \rightarrow & (R \vee (\Box_{\mathcal{T},k} R < S_0))^{I^*} \\ & & \rightarrow & R \vee \Box_{\mathcal{T},k}^{I^*} R \\ & & \rightarrow & R \end{array}$$

We take $\mathcal{P} := I\mathcal{M}$. Suppose, for any Σ_1^0-sentence S, that $\mathcal{U} \vdash S$. This tells us that $\mathcal{M} : U \rhd (T + S^I)$. Ergo $\mathcal{M} : U \rhd (T + R^{\mathcal{N}})$. We may conclude that $T \vdash R^{\mathcal{N}\mathcal{M}\mathcal{K}}$, i.o.w. $T \vdash R^{\mathcal{N}'}$. It now follows that

$$\begin{array}{rrcl} \mathcal{T} \vdash & \Box_{\mathcal{T},k}^J R & \rightarrow & R^{\mathcal{N}'} \wedge \Box_{\mathcal{T},k}^{J'} R \\ & & \rightarrow & R^{J'} \\ & & \rightarrow & R^J \\ & & \rightarrow & R \end{array}$$

Applying Löb's Rule, we have $\mathcal{T} \vdash R$. By cutelimination, we find $\mathcal{T} \vdash_k R$. Hence, by the external version of the proof of the FGH Theorem, we find that S_0 is true and, thus, that S is true. ❑

5 On the Manufacture of Faith

We repeat the definition of trustworthiness here.

Definition 5.1 A theory V is *trustworthy* if every U interpretable in V is also faithfully interpretable in V. ❑

In this section, we will provide a characterization of trustworthy theories. Friedman's result that consistent, finitely axiomatized, sequential theories are trustworthy, will follow from this characterization in combination with Theorem 4.1. Our treatment in this section can be viewed as generalizing some of Per Lindström's work on faithful interpretability. See [Lin97], chapter 6, §2. The methods used are for a great part those developed by Per Lindström and Viteslav Švejdar.

5.1 An Upper Bound

In this subsection we prove an upper bound result. We need two lemmas.

Lemma 5.2 Let $\mathcal{T} = \langle T, \mathcal{N} \rangle$ be a numerized theory. Let Γ be any class of T-sentences for which $\mathcal{T}$ contains a definable truth predicate, say TRUE. We only need that TRUE satisfies Tarski's convention. Suppose that the set of codes of elements of Γ has a fixed binumeration in $\mathcal{T}$. Then, there is a unary predicate of numbers $A(x)$, such that $\mathcal{T} \vdash (A(x) \wedge A(y)) \rightarrow x = y$, and such that, for any n, $\mathcal{T} + A(\underline{n})$ is Γ-conservative over $\mathcal{T}$. We may consider A as representing a closed partial numerical term τ, writing '$\tau \simeq x$' for '$A(x)$. ❑

We give the proof in appendix B.

Lemma 5.3 Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be a numerized theory. Let $\mathcal{L}$ be a language of finite signature σ for predicate logic. We call predicate logic of signature σ: FOL_σ. Let $\alpha(x)$ be any formula in the language of T such that $\mathcal{T}$ proves that all elements of $\{x \mid \alpha(x)\}$ are codes of $\mathcal{L}$-sentences. We write $\Box_\alpha$ for provability from the sentences coded by the elements of $\{x \mid \alpha(x)\}$. We write $\mathsf{con}(\alpha)$ for $\neg\, \Box_\alpha \bot$.

There is an interpretation $\mathcal{H} : (\mathcal{T} + \mathsf{con}(\alpha)) \rhd \mathsf{FOL}_\sigma$ such that, for any $\mathcal{L}$-sentence A, we have $\mathcal{T} + \mathsf{con}(\alpha) + \Box_\alpha A \vdash A^{\mathcal{H}}$. We say that $\mathcal{H}$ is a *Henkin interpretation* of α. ❑

Proof

We can see this by inspection of the usual proof of the Interpretation Existence Lemma. The basic idea is that we formalize the Henkin construction, employing definable cuts whenever we would have used induction in PA. See e.g. [Vis91] or [Vis92]. ❑

We proceed with our, somewhat technical, upperbound result. The bit with the sentence A is present, because we want our result to be applicable also to some theories that are *not* numerizable.

Lemma 5.4 Let T be any theory. Suppose $\mathcal{K} : T \rhd U$. Let A be any T-sentence. Suppose $\mathcal{W} = \langle T + A, \mathcal{N}\rangle$ is numerized. Then there is an interpretation $\mathcal{M} : T \rhd U$ such that, for any U-sentence B, $T \vdash B^{\mathcal{M}} \Rightarrow \mathcal{W} \vdash \Box_U B$. ❑

Proof

Consider $\mathcal{W}$. We can, by Fact 2.2(1) and Lemma 4.9, shorten $\mathcal{N}$ to a $\mathcal{W}$-definable cut J such that $\mathcal{Z} := \langle T + A, J\rangle$ contains a truth predicate for the Σ^0_1-sentences of $\mathcal{Z}$. (Remember that the meaning of 'Σ^0_1' shifts with the numerization.) Note that $\mathcal{Z} \vdash \Box_U B \Rightarrow \mathcal{W} \vdash \Box_U B$. It follows that it is sufficient to prove our theorem for $\mathcal{Z}$. Thus, we may, without loss of generality, assume that $\mathcal{W}$ contains a truth predicate, say true, for the Σ^0_1-sentences. Moreover, we may, by Fact 2.2(2), assume that $\mathcal{W}$ proves $I\Delta_0 + \Omega_1$.

Let τ be the partial closed term promised by Lemma 5.2 for $\mathcal{W}$ and Σ^0_1. We fix some standard enumeration C_x of the U-sentences in such a way that $\mathcal{W}$ verifies its elementary properties. We specify $\mathcal{M}$, in T, by cases. In case we have $\neg A$, we take $\mathcal{M}$ equal to $\mathcal{K}$. Suppose we have A. We may now work in $\mathcal{W}$. Let $U^* := U + \{C_x \mid \tau \simeq x\}$. Note that (i) U^* is not Δ^b_1-axiomatized, and that (ii) in talking about U^* we are really talking about the formula defining the axiom set and that (iii) the definition of U^* only makes sense in the presence of A. In case $\mathsf{incon}(U^*)$, we take $\mathcal{M}$ again equal to $\mathcal{K}$. If $\mathsf{con}(U^*)$, we take $\mathcal{M}$ equal to the Henkin-interpretation $\mathcal{H}$ of U^*. We give the clauses for $\mathcal{M}$, for the cases of the domain of the interpretation and the translation of a binary predicate:

- $\delta_{\mathcal{M}}(x) :\leftrightarrow ((\neg A \vee (A \wedge \mathsf{incon}^{\mathcal{N}}(U^*))) \wedge \delta_{\mathcal{K}}(x)) \vee (A \wedge \mathsf{con}^{\mathcal{N}}(U^*) \wedge \delta_{\mathcal{H}}(x))$,
- $P^{\mathcal{M}}(x,y) :\leftrightarrow ((\neg A \vee (A \wedge \mathsf{incon}^{\mathcal{N}}(U^*))) \wedge P^{\mathcal{K}}(x,y)) \vee$
  $(A \wedge \mathsf{con}^{\mathcal{N}}(U^*) \wedge P^{\mathcal{H}}(x,y))$.

(In writing e.g. '$\mathsf{incon}^{\mathcal{N}}(U^*)$', we intend no relativization of the formula defining the axiom set.)

Clearly, $\mathcal{M} : T \rhd U$. Suppose $T \vdash B^{\mathcal{M}}$. Let $\neg B = C_n$. We have:

$$\mathcal{W} + \tau \simeq \underline{n} \vdash \text{“}(U + \neg B) = U^*\text{”}.$$

Hence, $\mathcal{W} + (\tau = \underline{n}) + \mathsf{con}(U + \neg B) \vdash \neg B^{\mathcal{M}}$. Thus, $\mathcal{W} \vdash (\tau \simeq \underline{n}) \rightarrow \Box_U B$. By the Σ^0_1-conservativity of $\tau \simeq \underline{n}$, we find $\mathcal{W} \vdash \Box_U B$. ❑

5.2 The Characterization

In this subsection, we provide the promised characterization of trustworthiness and prove Friedman's result as a corollary.

Theorem 5.5 *Let T be any Δ^b_1-axiomatized theory. The following are equivalent.*

1. *T is trustworthy.*
2. *T has a (finite) extension which has a Σ^0_1-sound numerization.*
3. *T has a (finite) extension on which there is Σ^0_1-sound interpretation of Q.*
4. *There is a faithful interpretation of predicate logic with one binary relation symbol into T.*[10]

Proof

“(1) ⇒ (2)”. Suppose T is trustworthy. Say the (relational) signature of number theory is σ. Trivially, the predicate logic FOL_σ is interpretable in T. Hence, there is a faithful interpretation, say $\mathcal{K}$, of FOL_σ in T. It is easily seen that $\langle T + (\bigwedge \mathsf{F})^{\mathcal{K}}, \mathcal{K}\rangle$ is a Σ^0_1-sound numerization of an extension of T.

“(2) ⇒ (1)”. Suppose T has a (finite) extension which has a Σ^0_1-sound numerization, say $\mathcal{W}$. It follows, by Σ^0_1-soundness, that $\mathcal{W} \vdash \Box_U B$ implies $U \vdash B$.

Suppose $\mathcal{K} : T \rhd U$. By Lemma 5.4, we may conclude that there is a faithful interpretation $\mathcal{M} : T \rhd U$.

“(1) ⇒ (4)”. This is immediate.

[10] We might want to insist that predicate logic contains identity. In this case it is only necessary that the interpretation is faithful w.r.t. the fragment of the formulas containing only R.

"(4) $\Rightarrow$ (2)". Suppose $\mathcal{P}$ is a faithful interpretation of predicate logic with one binary relation symbol into T. There is a finitely axiomatized set theory, say S, in the language with just one binary relation symbol into which F is faithfully interpretable, say via $\mathcal{Q}$. See e.g. [MM94]. Hence, $\langle T + (\bigwedge \mathsf{S})^{\mathcal{P}}, \mathcal{Q}\mathcal{P}\rangle$ is a Σ^0_1-sound numerization of an extension of T.

"(3) $\Leftrightarrow$ (2)". This is immediate, by the fact that F can be interpreted in Q on a cut I. Cuts are downwards closed under $\leq$. So we can always convert a Σ^0_1 sound interpretation of Q into a Σ^0_1-sound interpretation of F. ❑

The definition of trustworthiness is 'neutral' w.r.t. arithmetical theories and the like, in that it does not mention the presence of any device allowing coding. It does not even mention specific signatures. Thus it is remarkable that a theory involving coding is connected via (2) of the theorem to trustworthiness. In appendix C, we will discuss a nice alternative formulation of (2) of the theorem. From Theorem 5.5 combined with Theorem 4.1, we may now immediately conclude to Friedman's Theorem.

Corollary 5.6 [**Friedman's Theorem**] Finitely axiomatized, sequential, consistent theories are trustworthy. ❑

Remark 5.7 We have proved Friedman's Theorem from Theorem 4.1. It is easily seen that, conversely, the existence of a Σ^0_1-sound cut again follows from Corollary 5.6. Consider a finitely axiomatized, numerized, sequential and consistent theory $\mathcal{T} = \langle T, \mathcal{N}\rangle$. By Corollary 5.6, there is a faithful interpretation $\mathcal{M}$ of F in T. Clearly, $\langle T, \mathcal{M}\rangle$ is Σ^0_1-sound. Ergo, by Fact 2.2(3) and the upwards persistence of Σ^0_1-sentences, we can find a $\mathcal{T}$-cut I such that $\langle T, I\rangle$ is Σ^0_1-sound. ❑

Example 5.8 $\mathsf{PA}+\mathsf{incon}(\mathsf{PA})$ is not trustworthy. This can be seen e.g. by noting that $\mathsf{PA} + \mathsf{incon}(\mathsf{PA}) \rhd \mathsf{PA}$. Since any interpretation of PA in $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$ is verifiably an end-extension of the identity interpretation, it will, by the upwards persistence of Σ^0_1-sentences, satisfy $\mathsf{incon}(\mathsf{PA})$. Hence no faithful interpretation of PA in $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$ is possible.

In contrast, $\mathsf{ACA}_0 + \mathsf{incon}(\mathsf{ACA}_0)$ is trustworthy. ❑

We may use Theorem 4.15 to get a modest strengthening of Friedman's Theorem.

Corollary 5.9 Suppose T is consistent, finitely axiomatized and sequential. Suppose T and U are mutually interpretable. Then U is trustworthy. ❑

Open Question 5.10 We could say that a theory T is *solid* if every U that is mutually interpretable with T is trustworthy. Is there a perspicuous characterization of solid theories?

Note that PA and $\mathsf{PA} + \mathsf{incon}(\mathsf{PA})$ are mutually interpretable. So, by Example 5.8, PA is trustworthy but not solid. ❑

We proceed with some further corollaries of Theorem 5.5. The following corollary of is easy.

Corollary 5.11 Any subtheory of a trustworthy theory is trustworthy. ❑

Corollary 5.12 Consider Group Theory group_c, where we allow an extra constant c in the language. The theory group_c is trustworthy. ❑

Proof

Tarski constructs, in [TMR53], a model $\mathcal{G}$ of group_c that has as definable inner model the natural numbers with plus and times. In other words, he constructs an interpretation $\mathcal{K}$ with $\mathcal{K} : \mathsf{Th}(\mathcal{G}) \rhd \mathsf{Th}(\mathbb{N})$. It follows that $\langle \mathsf{group}_c + (\bigwedge \mathsf{F})^{\mathcal{K}}, \mathcal{K} \rangle$ is Σ^0_1-sound. Ergo, by Theorem 5.5, group_c is trustworthy. ❑

Corollary 5.13 Any trustworthy theory is of degree $\mathbf{0}'$. ❑

Proof

This is immediate by Theorem 3.6. ❑

Open Question 5.14 What is the complexity of trustworthiness? Our characterization shows that this complexity is at most Σ^0_3. I conjecture that it is complete Σ^0_3. ❑

6 On the Nature of Trustworthiness

The notion of trustworthiness may, at first sight, seem to be somewhat artificial. Thus, one may wonder what structure is 'the natural home' of the notion. I am not sure this question has a unique answer. However, the answer given below is a good candidate. The answer will be that the relevant 'structure' is the embedding functor of two preorders.

Consider the preorder PFI of consistent theories ordered by the relation $\lhd_{\mathsf{f}}$, where $U \lhd_{\mathsf{f}} V$ if U is faithfully interpretable in V. We write $U \equiv_{\mathsf{f}} V$ for: $U \lhd_{\mathsf{f}} V$ and $V \lhd_{\mathsf{f}} U$. Consider also the preorder PI of consistent theories ordered by the relation $\lhd$, where $U \lhd V$ if U is interpretable in V. We write $U \equiv V$ for: $U \lhd V$ and $V \lhd U$.

These preorders can be viewed as categories in the usual way. If we divide out isomorphisms, we get the partial orderings of degrees of faithful interpretability and of degrees of interpretability.

Let emb be the identical embedding functor from PFI to PI. We will show that emb has a right adjoint, $\widetilde{(\cdot)}$, i.e. a mapping from theories to theories satisfying the magical equation[11]:

$$U \lhd_{\mathsf{f}} \widetilde{V} \Leftrightarrow \mathsf{emb}(U) \lhd V.$$

[11] See [Mac71], for the basic facts on adjunctions.

From this equation, the following facts are immediate consequences.

1. $\widetilde{(\cdot)}$ is a functor.
2. $\widetilde{V}$ is trustworthy.
3. $V \equiv \widetilde{V}$. So every degree of interpretability has a trustworthy element.
4. V is trustworthy iff $V \equiv_{\mathsf{f}} \widetilde{V}$.

We specify $\widetilde{(\cdot)}$. Consider a theory T. We expand the signature of T with a unary predicate P and with a binary predicate R. The theory $\widetilde{T}$ is the theory axiomatized by the axioms of T where we relativize the quantifiers to P. No non-logical principles concerning R are added. (The logical axioms concerning identity belong to predicate logic and are left unrelativized.) It is easily seen that (a) $T \equiv \widetilde{T}$. By a simple model-theoretical argument, we may show that $\widetilde{T}$ is conservative over predicate logic with just the binary relation symbol R. Hence, by translating R as $R(x, y)$, the theory $\widetilde{T}$ faithfully interprets predicate logic with just the binary relation symbol R. By Theorem 5.5(4), it follows that (b) $\widetilde{T}$ is trustworthy. From (a) and (b), it is immediate that $\widetilde{(\cdot)}$ is right adjoint to emb.

In case T has an infinite model, we can skip the relativization to P in the construction of $\widetilde{T}$. Thus we only need to expand the signature with R. Note that sequential theories are not closed under relativization of the domain. However, sequential theories are closed under adding predicate symbols. By the preceding observation, the mapping *add a binary relation symbol* will be right adjoint of the embedding functor, if we restrict both preorders to consistent sequential theories.

In case a numerization $\langle T, \mathcal{N} \rangle$ satisfies full induction, we can also take for $\widetilde{T}$, the theory $\mathsf{PA} + \{\mathsf{con}_n(T) \mid n \in \omega\}$, where $\mathsf{con}_n(T)$ means consistency of the set of axioms of T with Gödel number less than or equal to n. It follows that we can find an appropriate right adjoint, if we restrict both preorders to consistent extensions of PA in the arithmetical language.

By Theorem 5.9, consistent, finitely axiomatized, sequential theories T have the further property that if $T \equiv U$, then $T \equiv_{\mathsf{f}} U$. It is easy to see that this property is equivalent to the property of solidity introduced in Question 5.10.

References

[Bus86] S. Buss. *Bounded Arithmetic*. Bibliopolis, Napoli, 1986.

[Han65] W. Hanf. Model-theoretic methods in the study of elemtary logic. In J.W. Addison, L. Henkin, and A. Tarski, editors, *The Theory of Models, Proceedings of the 1963 International Symposium at Berkeley*, pages 132–145. North Holland, Amsterdam, 1965.

[HP91] P. Hájek and P. Pudlák. *Metamathematics of First-Order Arithmetic*. Perspectives in Mathematical Logic. Springer, Berlin, 1991.

[Kra87] J. Krajíček. A note on proofs of falsehood. *Archiv für Mathematische Logik und Grundlagenforschung*, 26:169–176, 1987.

[Lin97] P. Lindström. *Aspects of Incompleteness*, volume Lecture Notes in Logic 10. Springer, Berlin, 1997.

[Mac71] S. MacLane. *Categories for the Working Mathematician*. Number 5 in Graduate Texts in Mathematics. Springer, New York, 1971.

[MM94] F. Montagna and A. Mancini. A minimal predicative set theory. *The Notre Dame Journal of Formal Logic*, 35:186–203, 1994.

[Pud85] P. Pudlák. Cuts, consistency statements and interpretations. *The Journal of Symbolic Logic*, 50:423–441, 1985.

[Smo81] C. Smoryński. Fifty years of self-reference. *The Notre Dame Journal of Formal Logic*, 22:357–374, 1981.

[Smo85a] C. Smoryński. Nonstandard models and related developments. In L.A. Harrington, M.D. Morley, A. Šĕdrov, and S.G. Simpson, editors, *Harvey Friedman's Research on the Foundations of Mathematics*, pages 179–229. North Holland, Amsterdam, 1985.

[Smo85b] C. Smoryński. *Self-Reference and Modal Logic*. Universitext. Springer, New York, 1985.

[Šve83] V. Švejdar. Modal analysis of generalized rosser sentences. *The Journal of Symbolic Logic*, 48:986–999, 1983.

[TMR53] A. Tarski, A. Mostowski, and R.M. Robinson. *Undecidable theories*. North–Holland, Amsterdam, 1953.

[Vis91] A. Visser. The formalization of interpretability. *Studia Logica*, 51:81–105, 1991.

[Vis92] A. Visser. An inside view of exp. *The Journal of Symbolic Logic*, 57:131–165, 1992.

[Vis93] A. Visser. The unprovability of small inconsistency. *Archive for Mathematical Logic*, 32:275–298, 1993.

[Vis98] A. Visser. An Overview of Interpretability Logic. In M. Kracht, M. de Rijke, H. Wansing, and M. Zakharyaschev, editors, *Advances in Modal Logic, vol 1*, CSLI Lecture Notes, no. 87, pages 307–359. Center for the Study of Language and Information, Stanford, 1998.

[Vis99] A. Visser. Rules and Arithmetics. *The Notre Dame Journal of Formal Logic*, 40(1):116–140, 1999.

[VV94] L.C. Verbrugge and A. Visser. A small reflection principle for bounded arithmetic. *The Journal of Symbolic Logic*, 59:785–812, 1994.

[WP87] A. Wilkie and J.B. Paris. On the scheme of of induction for bounded arithmetic formulas. *Annals of Pure and Applied Logic*, 35:261–302, 1987.

[Yav97] R.E. Yavorsky. Logical schemes for first order theories. In *Springer LNCS (Yaroslavl'97 volume)*, volume 1234, pages 410–418, 1997.

A A Notational Convention

In this appendix we make the convention for the use of two kinds of variables and of boxes precise. Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$, $\mathcal{T}' = \langle T', \mathcal{N}\rangle$, ..., be numerized theories and let U, U', ..., be arbitrary theories. We assume that each theory comes equipped with a Δ_1^b-formula defining the axiom set. We treat the case, where we just have ordinary boxes in the language. Addition of e.g. $\Box_{\mathcal{T},n}$ and $\Box_{U,n}$ is entirely analogous.

We assume the language $\mathcal{L}_T$ of T has variables $\xi, \xi', \ldots$. We enrich $\mathcal{L}_T$ to a language $\mathcal{L}_\mathcal{T}$ with a second kind of variables $x, x', \ldots$ and with unary operators $\Box_U$ and $\Box_{\mathcal{T}'}$, for various U and $\mathcal{T}'$. The terms of the extended language are the smallest set containing both sets of variables and closed under the term-forming operations of $\mathcal{L}_T$. The set of formulas of $\mathcal{L}_\mathcal{T}$ is the smallest set F such that:

- $P(t_0, \cdots, t_{n-1})$ is in F, if the t_i are terms of the extended language and P is an n-ary predicate symbol of $\mathcal{L}_T$;
- F is closed under the propositional connectives and under the quantifiers $\forall x$, $\exists x$, $\forall \xi$, $\exists \xi$, for all variables x and ξ;
- If A is a sentence of $\mathcal{L}_U$, then $\Box_U A$ is in F;
- If A is a formula of $\mathcal{L}_{\mathcal{T}'}$ with only free variables in $x, x', y, \ldots$, then $\Box_{\mathcal{T}'} A$ is in F.

We can give the formulas of $\mathcal{L}_\mathcal{T}$ their desired translations into $\mathcal{L}_T$ via the translation $(\cdot)^\mathcal{T}$. We arrange it so that we have infinitely many variables $\eta, \eta', \ldots$ available in $\mathcal{L}_T$ distinct from the variables $\xi, \xi', \ldots$. We translate the terms by replacing x by η, x' by η', etcetera.

- $(P(t_0, \cdots, t_{n-1}))^\mathcal{T} := P(t_0^\mathcal{T}, \cdots, t_{n-1}^\mathcal{T})$;
- $(\cdot)^\mathcal{T}$ commutes with the propositional connectives and with $\forall \xi$, $\exists \xi$;
- $(\forall x\ A)^\mathcal{T} := \forall \eta\ (\delta_\mathcal{N}(\eta) \to A^\mathcal{T})$;
- $(\exists x\ A)^\mathcal{T} := \exists \eta\ (\delta_\mathcal{N}(\eta) \wedge A^\mathcal{T})$;
- $(\Box_U A)^\mathcal{T} := \mathsf{prov}_U(\underline{\#A})$;
- $(\Box_{\mathcal{T}'} A)^\mathcal{T} := \mathsf{prov}_{T'}(\underline{\#(A^{\mathcal{T}'})})$.
  (The numerical variables in A are treated in the usual way.)

B Conservativity

In this appendix, we prove Lemma 5.2. Let $\mathcal{T} = \langle T, \mathcal{N} \rangle$ be a numerized theory. Let Γ be any class of T-sentences for which $\mathcal{T}$ contains a definable truth predicate, say TRUE. We only need that TRUE satisfies Tarski's convention. We assume that the set of codes of elements of Γ has a fixed binumeration in T. We show that there is a unary predicate of numbers $A(x)$, such that $\mathcal{T} \vdash (A(x) \wedge A(y)) \rightarrow x = y$, such that, for any n, $\mathcal{T} + A(\underline{n})$ is Γ-conservative over $\mathcal{T}$.

We define, in $\mathcal{T}$, using the Gödel Fixed Point Lemma, the formula $A(x)$ as follows.

$$\begin{aligned} A(x) \quad \leftrightarrow \quad & \exists p \, \exists C{\in}\Gamma \, (\, \mathsf{proof}_{\mathcal{T}}(p, A(x) \rightarrow C) \wedge \neg \, \mathsf{TRUE}(C) \wedge \\ & \forall q < p \, \forall D{\in}\Gamma \, \forall y \, (\, \mathsf{proof}_{\mathcal{T}}(q, A(y) \rightarrow D) \rightarrow \mathsf{TRUE}(D) \,) \,) \end{aligned}$$

We assume that the formalization of proof is standard, so that every proof has a single conclusion C with $C < p$, etc. We first prove the uniqueness clause. Reason in $\mathcal{T}$. Suppose that $x \neq y$ and $A(x)$ and $A(y)$. Let p be a witness for $A(x)$ and let q be a witness of $A(y)$. By our assumption about the proof predicate, it follows that $p \neq q$. since in F, we have the linearity of $<$, it follows that $p < q$ or $q < p$. By the specification of A, it follows that this is impossible.

We move to the metatheory again. We prove our theorem by induction on $\mathcal{T}$-proofs. Suppose, that for all T-proofs $q < p$, we have, if $q : \mathcal{T} \vdash A(\underline{m}) \rightarrow D$, for some m and for some $D \in \Gamma$, then $\mathcal{T} \vdash D$. ('$r : \mathcal{T} \vdash E$' means: r is a $\mathcal{T}$-proof of E.) Suppose further that $p : \mathcal{T} \vdash A(\underline{n}) \rightarrow C$, for $C \in \Gamma$. We show $\mathcal{T} \vdash C$. From our assumptions, we have the following propositions.

$$\mathcal{T} \vdash (C \in \Gamma) \tag{1}$$

$$p : \mathcal{T} \vdash A(\underline{n}) \rightarrow C \tag{2}$$

It follows that:

$$\mathcal{T} + \neg \, C \vdash \neg A(\underline{n}) \tag{3}$$

$$\mathcal{T} + \neg \, C \vdash C \in \Gamma \wedge \mathsf{proof}_{\mathcal{T}}(\underline{p}, A(\underline{n}) \rightarrow C) \wedge \neg \, \mathsf{TRUE}(C) \tag{4}$$

Using (3), (4) and the specification of A, we may conclude that:

$$\mathcal{T} + \neg \, C \vdash \exists q < \underline{p} \, \exists D{\in}\Gamma \, \exists y \, (\, \mathsf{proof}_{\mathcal{T}}(q, A(y) \rightarrow D) \wedge \neg \, \mathsf{TRUE}(D) \,) \tag{5}$$

It follows that:

$$\mathcal{T} + \neg \, C \vdash \bigvee_{q<p, D<p, D\in\Gamma, m<p} (\, \mathsf{proof}_{\mathcal{T}}(\underline{q}, A(\underline{m}) \rightarrow D) \wedge \neg \, D \,) \tag{6}$$

Consider any $q < p$, $D < p$ with $D \in \Gamma$, and $m < p$. In case we have: $q : \mathcal{T} \vdash A(\underline{m}) \rightarrow D$, it follows, by the minimality of p, that $\mathcal{T} \vdash D$. In this

case the disjunct corresponding to q in (6) is $\mathcal{T}$-provably equivalent to absurdity and may be omitted. Suppose that q does not witness $\mathcal{T} \vdash A(\underline{m}) \to D$, then, by Σ-completeness, $\mathcal{T} \vdash \neg\, \mathsf{Proof}_{\mathcal{T}}(\underline{q}, A(\underline{m}) \to D)$. So again we may omit the disjunct corresponding to q. Thus the whole disjunction of (6) reduces to $\bot$. We may conclude: $\mathcal{T} \vdash C$. Quod erat demonstrandum.

Remark B.1 Let's assume that Γ is closed under disjunction. Let $\mathcal{W}$ be the theory axiomatized by the axioms of $\mathcal{T}$, plus the negations of false Γ-sentences. We use the obvious formula for the axiom set of $\mathcal{W}$ in $\mathcal{T}$. We write $\Box^*$ for provability in $\mathcal{W}$. Suppose B is of the form $\exists y\, B_0$. We write $\exists\!\!\exists x\, B$, for $\exists y \exists x\, B_0$. Under these conventions we can rewrite the specification of A as follows.

$$\mathcal{T} \vdash A(x) \leftrightarrow \Box^* \neg A(x) \leq \exists\!\!\exists y\, \Box^* \neg A(y).$$

It would be interesting to see a modal treatment of our argument. ❑

C Derivable Consequence

In this appendix, we provide reformulations of some of our results in terms of derivable consequence. Let T be a theory and let τ be a signature. We define some consequence relations for signature τ. Let Γ and Δ be sets of sentences of the language of signature τ and let A be a sentence of the language of signature τ.

- $\Gamma \mid \Delta \Vdash^*_T A :\Leftrightarrow \forall \mathcal{K}\, (T \vdash \Gamma^{\mathcal{K}} \Rightarrow T + \Delta^{\mathcal{K}} \vdash A^{\mathcal{K}})$.
- $\Delta \Vdash_T A :\Leftrightarrow \emptyset \mid \Delta \Vdash^*_T A$.
- $\Gamma \ \sim_T A :\Leftrightarrow \Gamma \mid \emptyset \Vdash^*_T A$.
- $\Lambda^\tau_T := \{A \mid \emptyset \Vdash_T A\}$.

Here it is implicitely assumed that the '$\mathcal{K}$' are interpretations for τ. (If τ is not clear from the context, we will exhibit it as superscript.) $\Vdash$ is the relation of *T-derivable consequence* and $\sim$ is the relation of *T-admissible consequence*. Λ^τ_T is the predicate logic of T (for signature τ). For some remarks on admissible consequence, see [Vis99] appendix A. For some information about derivable consequence, see [Vis98], subsection 12.3. For a study of predicate logics of classical theories, see [Yav97]. Here are some elementary facts about these notions.

Fact C.1

1. $\Gamma \Vdash_T A \Leftrightarrow \forall U \supseteq T\ \Gamma \ \sim_U A$.
   Here 'U' ranges over theories with arbitrarily complex axiom sets.
2. $A \Vdash^\tau_T B \Leftrightarrow \Lambda^\tau_T \vdash (A \to B)$.
3. If $T \rhd^\tau A$ and $A \ \sim^\tau_T B$, then $A \Vdash^\tau_T B$.
4. We can find T, A, B, such that $A \ \sim^\tau_T B$, but not $A \Vdash^\tau_T B$.

❑

Proof

Ad (3). Suppose $\mathcal{K} : T \rhd^\tau A$ and $A \ \vdash^\tau_T B$. Consider any interpretation $\mathcal{M}$ for τ. We construct a new interpretation $\mathcal{P}$ as follows: $\mathcal{P}$ is $\mathcal{K}$ if $\neg A^{\mathcal{M}}$ and $\mathcal{P}$ is $\mathcal{M}$ if $A^{\mathcal{M}}$. Clearly, $\mathcal{P} : T \rhd^\tau A$. By $A \ \vdash^\tau_T B$, it follows that $\mathcal{P} : T \rhd^\tau B$. Ergo: $T \vdash (A^{\mathcal{M}} \to B^{\mathcal{M}})$.

Ad (4). Let σ be the signature of arithmetic. We have $\bigwedge \mathsf{F} \wedge \mathsf{con}(T) \ \vdash^\sigma_T \bot$, for any T. This is, in fact, Pudlák's strong version of the Second Incompleteness Theorem. On the other hand, if we take T e.g. PA, clearly $\bigwedge \mathsf{F} \wedge \mathsf{con}(T) \ \nvdash^\sigma_T \bot$. ❑

A T-model $\mathcal{N}$ of signature τ, is a model for signature τ that is isomorphic to an internal model of a model $\mathcal{M}$ of T. Internal models are given by interpretations $\mathcal{K}$. We could call the internal model of $\mathcal{M}$ given by $\mathcal{K}$: $\mathcal{K}^{\mathcal{M}}$. Thus, $\mathcal{N}$ is a T-model iff, for some model $\mathcal{M} \models T$ and for some interpretation for signature τ, $\mathcal{N}$ is isomorphic to $\mathcal{K}^{\mathcal{M}}$. We can understand $\Vdash$ in terms of T-models, as follows.

Fact C.2 $\Delta \Vdash_T A :\Leftrightarrow$ for all T-models $\mathcal{N}$, $(\mathcal{N} \models \Delta \Rightarrow \mathcal{N} \models A)$. ❑

Here is an example illustrating the non-compactness of $\vdash$.

Example C.3 Let σ be the signature of arithmetic. Let T be a finitely axiomatized, consistent, sequential theory. Let $U := \mathsf{F} + \{\mathsf{con}_n(T) \mid n \in \omega\}$. (Here, we can use either the complexity measure 'depth of connectives' or the measure 'depth of quantifier changes'.) Then, since U is locally, but not globally interpretable in T, we find that $\vdash$ is not compact. ❑

To provide an example to illustrate the non-compactness of $\Vdash$, we need a result of Jan Krajíček.

Theorem C.4 *Let $\mathcal{T} = \langle T, \mathcal{N} \rangle$ be $I\Delta_0$ or let $\mathcal{T}$ be finitely axiomatized, consistent and sequential. There is a mapping $I \mapsto k_I$, from $\mathcal{T}$-cuts to natural numbers, such that the theory*

$$\mathsf{kraj}(\mathcal{T}) := \mathcal{T} + \{\mathsf{incon}^I_{k_I}(T) \mid I \text{ is a } T\text{-cut}\}$$

is locally interpretable in $\mathcal{T}$, and, hence, consistent.[12] *Here the complexity measure used is depth of connectives. We can, however, also use depth of quantifier changes.*

For a proof, see [Kra87], section 3. The functionality suggested by our notation '$\mathsf{kraj}(\mathcal{T})$' is par abus de langage, since the theory does not seem to be uniquely determined by the data. In fact, I have the following conjecture.

[12] By inspecting the argument, it becomes clear that Krajíček's theory is recursively enumerable. I did not check that the axioms are indeed p-time decidable. However, we can always apply Craig's trick to obtain a p-time decidable axiomatization. Note that the verification of Craig's trick demands a metatheory containing Σ^0_1-collection.

Open Question C.5 Prove or refute the following conjecture. There are infinitely many theories satisfying the description of $\mathsf{kraj}(\mathcal{T})$ that are pairwise not mutually interpretable. ❑

By construction, the theory $\mathsf{kraj}(\mathcal{T})$ is not trustworthy. It follows from Theorem 5.9 that $\mathsf{kraj}(\mathcal{T})$ is *not* globally interpretable in $\mathcal{T}$. We can now present the promised example for the non-compactness of $\Vdash$.

Example C.6 Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be $I\Delta_0$ or let $\mathcal{T}$ be finitely axiomatized, consistent and sequential. Let $U := \mathsf{F} + \{\mathsf{con}_n(T) \mid n \in \omega\}$. Now it is easy to see that $U \Vdash_{\mathsf{kraj}(\mathcal{T})} \bot$, but that, for no finite subtheory U_0 of U, we have $U_0 \Vdash_{\mathsf{kraj}(\mathcal{T})} \bot$. Hence, $\Vdash$ is not compact. ❑

Our notions have at most complexity Π^0_2. The following theorem shows that the worst may happen.

Theorem C.7 *There is a theory W such that Λ^σ_W is complete Π^0_2. Here σ is the signature of arithmetic. It follows that $\Vdash^*$, $\Vdash$, $\mathrel{|\!\sim}$ and Λ assume their maximal possible complexities*

Proof

Let $\mathcal{T} = \langle T, \mathcal{N}\rangle$ be $I\Delta_0$ or finitely axiomatized, consistent and sequential. Let $\mathcal{W} = \langle W, \mathcal{N}\rangle := \mathsf{kraj}(\mathcal{T})$ as inTheorem C.4. We show that Λ^σ_W is complete Π^0_2.

Consider the sentence $A := \forall x\, \exists y\ A_0(x, y)$, where $A_0 \in \Delta_0$. Let $S_x := \exists y\, A_0(x, y)$. Let R_x the FGH-sentence for $\mathcal{W}$ and S_x. We define:

$$Q := \forall x\ (\mathsf{con}_x(T) \to R_x).$$

We show that A iff $\Lambda^\sigma_W \vdash \bigwedge \mathsf{F} \to Q$.

Suppose that A. Let $\mathcal{K}$ be any interpretation for the signature σ. Consider the interpretation $\mathcal{M}$ such that, in T, $\mathcal{M}$ is $\mathcal{K}$ if $(\bigwedge \mathsf{F})^{\mathcal{K}}$ and $\mathcal{N}$, otherwise. Clearly, $\mathcal{M} : T \rhd \mathsf{F}$. By Fact 2.2(3), there is a T-cut I of $\mathcal{N}$, such that $I \leq_T \mathcal{M}$. By the construction of W, we have $\Box_W \mathsf{incon}^I_n(T)$, for some n. It follows that $\Box_W \mathsf{incon}^{\mathcal{M}}_n(T)$. We may conclude that $\Box_W((\bigwedge \mathsf{F})^{\mathcal{K}} \to \mathsf{incon}^{\mathcal{K}}_n(T))$. It follows that (a):

$$\Box_W\, ((\bigwedge \mathsf{F})^{\mathcal{K}} \to (\forall x\ (\mathsf{con}_x(T) \to x < n))^{\mathcal{K}}).$$

From A we can infer, for any k, that S_k. Hence $R_k \vee R^\bot_k$. From R_k, we have, by Σ^0_1-completeness, $\Box_W((\bigwedge \mathsf{F})^{\mathcal{K}} \to R^{\mathcal{K}}_k)$. In case, $R^\bot_k$, we have $\Box_{\mathcal{W}} R_k$, by the definition of R_k, and $\Box_{\mathcal{W}} R^\bot_k$, by Σ^0_1-completeness. Hence $\Box_{\mathcal{W}} \bot$, quod non. Ergo we find: $\Box_W((\bigwedge \mathsf{F})^{\mathcal{K}} \to (\bigwedge_{k<n} R_k)^{\mathcal{K}})$. Thus, we get (b):

$$\Box_W((\bigwedge \mathsf{F})^{\mathcal{K}} \to (\forall x < n\ R_x)^{\mathcal{K}}).$$

We may conclude, combining (a) and (b), that $\Box_W((\bigwedge \mathsf{F})^{\mathcal{K}} \to Q^{\mathcal{K}})$. Thus, $\Lambda^{\sigma}_W \vdash \bigwedge \mathsf{F} \to Q$.

For the converse, suppose that $\Lambda^{\sigma}_W \vdash \bigwedge \mathsf{F} \to Q$. Consider any n. Pick a $\mathcal{T}$-cut I such that $\Box_W \mathsf{con}^I_n(T)$. By our assumption, we have $\Box_W Q^I$. Hence, $\Box_W R^I_n$ and, so, $\Box_{\mathcal{W}} R_n$. By the FGH-Theorem, we may conclude that S_n. ❑

We show how various notions of this paper can be formulated in a natural way in terms of derivable and admissible consequence. Let σ be the signature of arithmetic. We need the following definitions.

- Let T_n be $\{0 = 0\}$ if $n = 0$ and the set of true Π^0_n-sentences otherwise.
- Suppose $\mathsf{F} \subseteq \Gamma$. We define: $\Gamma \Vdash^n_T A :\Leftrightarrow \Gamma, \mathsf{T}_n \Vdash_T A$.
- Suppose $\mathsf{F} \subseteq \Gamma$. We define: $\Gamma \ \sim^n_T A :\Leftrightarrow \Gamma \mid \mathsf{T}_n \Vdash^*_T A$.
- $T, n\text{-}\mathsf{con}_{\mathsf{adm}}(U)$ iff not $U \ \sim^n_T \bot$.
- $T, n\text{-}\mathsf{con}(U)$ iff not $U \Vdash^n_T \bot$.

We now have:

Fact C.8 1. $T, n\text{-}\mathsf{con}_{\mathsf{adm}}(U)$ implies $T, n\text{-}\mathsf{con}(U)$.

2. $T, n\text{-}\mathsf{con}(U)$ iff there is a Σ^0_n-sound T-model of U.
3. A theory T is consistent and numerizable iff $T, 0\text{-}\mathsf{con}_{\mathsf{adm}}(\mathsf{F})$.
4. If $T, 0\text{-}\mathsf{con}(\mathsf{Q})$, then T is undecidable.

   (This follows from Tarski's Theorem that if an essentially undecidable theory is interpretable in a consistent extension of a given theory T, then T is undecidable. In fact $T, 0\text{-}\mathsf{con}(U)$ iff U is weakly interpretable in T.)
5. Let T be finitely axiomatized, consistent and sequential. Then, we have $T, 1\text{-}\mathsf{con}_{\mathsf{adm}}(\mathsf{Q})$. (This follows from Theorem 4.1.)
6. T is trustworthy iff $T, 1\text{-}\mathsf{con}(\mathsf{Q})$. (This follows from Theorem 5.5.)

   We may conclude that $T, 1\text{-}\mathsf{con}(\mathsf{Q})$ implies $T, n\text{-}\mathsf{con}(\mathsf{Q})$, for all n.

Note that Q in the above statements can be replaced by F or S^1_2 or $I\Delta_0$, by the fact that these stronger theories are interpretable on a cut in Q. ❑

Remark C.9 Consider a Δ^b_1-axiomatizable theory T satisfying $T, 0\text{-}\mathsf{con}(\mathsf{Q})$. By Theorem 3.6, T is in Turing degree $\mathbf{0}'$. William Hanf showed that there are finitely axiomatized T in any recursively enumerable Turing degree. (Even that there are essentially undecidable, finitely axiomatized theories of any recursively enumerable degree of unsolvability.) See [Han65]. Ergo, there are finitely axiomatized, undecidable theories T such that $T, 0\text{-}\mathsf{incon}(\mathsf{Q})$. ❑

D On the Existential Axioms of Q

In this appendix, we discuss a detail of the proof of Wilkie's Theorem that $I\Delta_0$ is interpretable on an initial segment in Q. *An initial segment*, is a definable set of numbers Q-provably closed under S and downwards closed under $\leq$.[13]

Our presentation is directly dependent on the presentation of Petr Hájek and Pavel Pudlák in their book [HP91] on pp. 369, 370. The reader is advised to first look at Hájek and Pudlák's proof. The axioms of Q are the following.

$\mathsf{Q}1 \vdash \mathsf{S}x \neq 0$,

$\mathsf{Q}2 \vdash \mathsf{S}x = \mathsf{S}y \rightarrow x = y$,

$\mathsf{Q}3 \vdash x \neq 0 \rightarrow \exists y\, x = \mathsf{S}y$,

$\mathsf{Q}4 \vdash x + 0 = x$,

$\mathsf{Q}5 \vdash x + \mathsf{S}y = \mathsf{S}(x + y)$,

$\mathsf{Q}6 \vdash x \times 0 = 0$,

$\mathsf{Q}7 \vdash x \times \mathsf{S}y = (x \times y) + x$,

$\mathsf{Q}8 \vdash x \leq y \leftrightarrow \exists z\, z + x = y$.

To prove Wilkie's Theorem, it is convenient to take $\leq$ a primitive symbol. If we would take it as defined by $\exists z\, z+x = y$, then we would have to state explicitely that on an initial segment I, the meaning of $\leq$ is preserved, i.e. that $\leq^I$ is equal to $\leq \restriction I$. This sameness of meaning is important, since we want downwards preservation of Π^0_1-sentences to the initial segment and upwards preservation of Σ^0_1-sentences from the segment. These preservation results are e.g. used to get initial segments with more and more Δ_0-induction.

Hájek and Pudlák wisely choose to treat $\leq$ as a primitive symbol. However, on p369, in their proof of Wilkie's Theorem, they stumble in the last step. They write: "... we can trivially interpret Q by eliminating $\leq$ from the language and deleting Q8." In other words, they redefine $\leq$. This argument won't wash, since they need the new $\leq$ on the initial segment to be the restriction of the old $\leq$ to the inital segment. Otherwise, the central argument does not go through.

Fortunately the gap in the argument of Hájek and Pudlák is easily closed by proceeding analogously to their verification of Q3 on the initial segment I: prove, by induction on x, that $\forall y{\leq}\, x\, \exists z{\leq} x\, z + y = x$. (We need some auxiliary inductions to show e.g. that $x \leq x$ and $x \leq \mathsf{S}x$.)

However, the problem to verify the existential axioms Q3 and Q8 also occurs in the case of the interpretation of Hájek and Pudlák's theory Q^+ in Q. For this reason, I prefer another strategy to settle the problem of these axioms for once and for all, right from the start.

[13] Our *initial segment* is Hájek and Pudlák's *cut*.

We work in Q. We use the easily verifiable theorem that $x+y=0 \rightarrow x=y=0$. A number x is *L-successive* iff $\forall y \forall z\ (y+z=x \rightarrow \mathsf{S}y+z=\mathsf{S}x)$. We show that 0 is L-successive. Suppose that $y+z=0$. Then $y=z=0$. Moreover, $\mathsf{S}y+z=\mathsf{S}0+0=\mathsf{S}0$. Next we show that the L-successive numbers are closed under successor. Suppose x is L-successive and suppose $y+z=\mathsf{S}x$. We want to show that $\mathsf{S}y+z=\mathsf{SS}x$. In case $z=0$, we have $y=\mathsf{S}x$ and, hence, $\mathsf{S}y+0=\mathsf{S}y=\mathsf{SS}x$. In case $y=\mathsf{S}u$, we have $y+\mathsf{S}u=\mathsf{S}x$. So, $\mathsf{S}(y+u)=\mathsf{S}x$. Ergo $y+u=x$. Since x is L-successive, we have $\mathsf{S}y+u=\mathsf{S}x$ and, so, $\mathsf{S}(\mathsf{S}y+u)=\mathsf{SS}x$. We may conclude that $\mathsf{S}y+\mathsf{S}u=\mathsf{SS}x$.

A number x is a *commutator* iff $\forall y \forall z\ (y+z=x \rightarrow z+y=x)$. We say that x is a *strong commutator* iff x is L-successive and x is a commutator. We already know that 0 is L-successive. Moreover, if $y+z=0$, then $y=z=0$, and, hence, $z+y=0$. So 0 is a strong commutator. We show that the strong commutators are closed under successor. Suppose x is a strong commutator. By the above argument, $\mathsf{S}x$ is L-successive. Suppose $y+z=\mathsf{S}x$. To show: $z+y=\mathsf{S}x$. First suppose $z=0$. We have $y=y+0=\mathsf{S}x$. So we need to show that $0+\mathsf{S}x=\mathsf{S}x$. We have $x+0=x$. So, since x is a commutator, we find $0+x=x$ and, hence $0+\mathsf{S}x=\mathsf{S}x$.. Next, suppose $z=\mathsf{S}u$. We have $y+\mathsf{S}u=\mathsf{S}x$. Then, $y+u=x$. Hence, since x is a commutator, we have $u+y=x$. Ergo, since x is L-successive, $\mathsf{S}u+y=\mathsf{S}x$.

Theorem D.1 (in Q) *Suppose, that the elements of an initial segment I are all commutators. Then, the segment I verifies* $\mathsf{Q}3$ *and* $\mathsf{Q}8$.

Proof

Reason in Q. Suppose $\mathsf{S}x$ is in I. We have $x+\mathsf{S}0=\mathsf{S}x$. Since, $\mathsf{S}x$ is a commutator, we find $\mathsf{S}0+x=\mathsf{S}x$. Ergo $x \leq \mathsf{S}x$. Hence $x \in I$. Thus any non-zero number in I has a predecessor in I.

Suppose $x \leq y$, for $y \in I$. Then, for some z, $z+x=y$. Since, y is a commutator, we find $x+z=y$, and so $z \leq y$. Hence $z \in I$. ❑

Now we execute the remaining part of the proof of Wilkie's Theorem inside the strong commutators, without worries about $\mathsf{Q}3$ and $\mathsf{Q}8$. We need closure of the strong commutors under successor to construct the appropriate initial segments.