# A NOTE ON THE STRENGTH OF DISENTANGLED TRUTH-THEORIES

RICHARD KIMBERLY HECK

ABSTRACT. So-called 'disentangled' truth-theories are supposed to prevent assumptions about the truth of statements in the object-language from inadvertently strengthening the background syntax. In earlier work, I proved some limitative results in an attempt to show that the strategy works, but those results leave several questions unanswered. We address some of them here. We also discuss a subtlety that has so far been overlooked in discussions of these theories.

Every axiom of Peano Arithmetic is true. The rules of inference preserve truth. So every theorem of PA is true. Since $0 = 1$ is not true, it is not a theorem of PA. So PA is consistent.

Such 'soundness arguments' play an important role in discussion of theories of truth. Because they establish the consistency of the 'base theory'—PA in this case—any theory in which such an argument can be formalized must be stronger than the base theory, by Gödel's second incompleteness theorem. But there are some puzzling aspects about the way such arguments are usually formalized.

We start with some arithmetical base theory $\Sigma$; we assume that $\Sigma$ is strong enough to be able to talk about its own syntax, via Gödel numbering. We then expand the language by adding semantic vocabulary—a truth-predicate and the like—and extend $\Sigma$ with Tarski-style compositional axioms. Those axioms will allow us to prove the 'T-sentences' for the language of arithmetic

$$A \equiv \mathsf{Tr}(\ulcorner A \urcorner)$$

and so to prove that each axiom of $\Sigma$ is itself true, by a trivial argument:

(1) $A$, since $A$ is an axiom
(2) $A \equiv \mathsf{Tr}(\ulcorner A \urcorner)$, since the T-sentences are provable
(3) So, $\mathsf{Tr}(\ulcorner A \urcorner)$

To carry out the induction that will take us from the truth of the axioms to the truth of the theorems, we also need to extend whatever induction axioms are present in $\Sigma$ to allow semantic vocabulary to occur in those axioms: The most natural formalization requires $\Sigma$ to contain induction for $\Sigma_1$ formulas (Heck, 2015, Theorem 3.20), though it is in fact sufficient

to have induction for $\Delta_0$ formulas (Łełyk, 2022).[1] Call the result of extending $\Sigma$ in this way $\mathsf{CT}[\Sigma]$.[2] Then, so long as $\Sigma$ extends $\mathsf{I}\Delta_0$,[3] $\mathsf{CT}[\Sigma]$ will prove $\mathsf{Con}(\Sigma)$.

This works, but it is in some ways unsatisfying. Suppose we want in this way to prove $\mathsf{Con}(\mathsf{PA})$. So we work in $\mathsf{CT}[\mathsf{PA}]$. Only finitely many of PA's axioms can be used in the proof (and PA is not finitely axiomatizable). So $\mathsf{Con}(\mathsf{PA})$ must be provable in $\mathsf{CT}[\mathsf{I}\Sigma_n]$, for some $n$. Moreover, there's a lacuna in the proof sketched earlier. It's clear enough how $\mathsf{CT}[\mathsf{PA}]$ proves that *each* axiom of PA is true. But how can we prove the general statement that *all* axioms of PA are true?[4] The answer is: by induction. There is, in fact, a single $\Sigma_1$ formula, in the extended language, from which all the infinitely many induction axioms of PA follow (Heck, 2015, pp. 447–50). What this means is that $\mathsf{CT}[\mathsf{I}\Sigma_1]$ contains PA. And, as noted, $\Sigma_1$ induction is enough to carry out the induction at the heart of the soundess argument. So, in fact, $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves $\mathsf{Con}(\mathsf{PA})$ already (Heck, 2015, Theorem 3.20), and in fact $\mathsf{CT}[\mathsf{I}\Delta_0]$ does so (Łełyk, 2022).

You might have thought that $\mathsf{CT}[\mathsf{I}\Sigma_1]$ would prove $\mathsf{Con}(\mathsf{I}\Sigma_1)$, $\mathsf{CT}[\mathsf{I}\Sigma_2]$ would prove $\mathsf{Con}(\mathsf{I}\Sigma_2)$, and so forth, and that we'd need $\mathsf{CT}[\mathsf{PA}]$ to prove $\mathsf{Con}(\mathsf{PA})$. But no. If you try to prove $\mathsf{Con}(\mathsf{I}\Sigma_1)$ via this kind of argument, you end up doing so in a theory already capable of proving $\mathsf{Con}(\mathsf{PA})$. That's what's unsatisfying.

The reason we get this result is that the 'base theory' is playing two different roles. Suppose we naïvely pursue the mentioned strategy in an attempt to prove $\mathsf{Con}(\mathsf{I}\Sigma_1)$. So we work in $\mathsf{CT}[\mathsf{I}\Sigma_1]$. Then, on the one hand, $\mathsf{I}\Sigma_1$ is the theory whose consistency we are proving. In that role, it is what allows us to prove the truth of $\mathsf{I}\Sigma_1$'s axioms, via the trivial argument mentioned above.[5] On the other hand, an extension of $\mathsf{I}\Sigma_1$ is also the theory *in which* we are proving consistency: It's what allows us to reason about syntax and semantics. But when we extend the induction axioms of $\mathsf{I}\Sigma_1$ *qua* syntax and semantics—which we need to do in order to carry out the induction that takes us from the truth of the axioms to

[1] A formula is $\Delta_0$ if all quantifiers occurring in it are bounded, i.e., are of the form $\forall x < t$ or $\exists x < t$; $\Sigma_1$if it is of the form $\exists v_1 \dots \exists v_n(\phi)$, where $\phi$ is $\Delta_0$; $\Pi_1$, of the form $\forall v_1 \dots \forall v_n(\phi)$, where $\phi$ is $\Delta_0$; $\Sigma_n$, of the form $\exists v_1 \dots \exists v_n(\phi)$, where $\phi$ is $\Pi_{n-1}$; $\Pi_n$, of the form $\forall v_1 \dots \forall v_n(\phi)$, where $\phi$ is $\Sigma_{n-1}$.

[2] We'll not be interested here in the weaker theory $\mathsf{CT}^-[\Sigma]$, in which that has not been done. As is well known, $\mathsf{CT}^-[\Sigma]$ is a conservative extension of $\Sigma$ and so does not prove $\mathsf{Con}(\Sigma)$. That does not mean, however, that $\mathsf{CT}^-[\Sigma]$ is no stronger than $\Sigma$ itself. In fact, if $\Sigma$ is finitely axiomatizable, it is stronger (Heck, 2015, §3.2).

[3] $\mathsf{I}\Delta_0$ is Robinson arithmetic, Q,plus the induction axioms for $\Delta_0$ formulas. Similarly, for $\mathsf{I}\Sigma_n$.

[4] Compare: For each $n$, PA proves that $n$ does not code a proof of $0 = 1$. But PA does not prove the general statement: For all $n$, $n$ does not code a proof of $0 = 1$, since that just is $\mathsf{Con}(\mathsf{PA})$.

[5] $\mathsf{I}\Sigma_1$ is finitely axiomatizable. So, if we use some finite axiomatization, we do not have to worry about how to get from 'each' to 'all'.

the truth of the theorems—we simultaneously give ourselves the ability to prove the truth of *much more* than just the axioms of $\mathsf{I}\Sigma_1$. In fact, we are now able to prove the truth of all the axioms of PA. The semantic induction axioms, as it were, inadvertently strengthen the base theory itself.

In earlier work (Heck, 2009, 2015), suggested that we might address this problem by following Tarski (1956) and 'disentangling' the two roles played by the base theory.[6] We now work in a multi-sorted framework. One sort contains numbers: the objects that our 'target' theory is about.[7] Another sort contains syntactic objects—terms, formulas, proofs, and the like, from the language of the target theory—and we have a syntactic theory that allows us to reason about these objects. It might, for example, contain axioms that allow us to carry out induction on the complexity of expressions, or on the length of proofs. A third sort contains assignments: functions from variables to their values. We then have a semantic theory that relates the syntactic objects to the numbers. Officially, we might think of the syntactic theory as formulated in a language whose sole primitive is a symbol for concatenation: $x \frown y$, so that the theory really is a theory of syntax.[8] For convenience and familiarity, however, we may take the syntactic theory also to be an arithmetical theory. Still, we want to think of the syntactic language $\mathscr{S}$ as distinct from the language of arithmetic, $\mathscr{A}$. Perhaps it is "written in boldface, or something of the sort" (Heck, 2015, p. 451).

Let $\mathsf{CTD}[\Sigma]$ be the semantic theory for the language of arithmetic built on the syntactic theory $\Sigma$.[9] As before, a natural formalization of the soundness argument requries $\Sigma_1$ induction, so we assume that $\Sigma$ contains $\mathsf{I}\Sigma_1$. If we want to prove the consistency of some arithmetical theory $\Theta$ in this way, then we also need to assume that $\Theta$'s axioms are true. But, if we do, then we can indeed prove $\mathsf{Con}(\Theta)$. That is:[10]

**Theorem 1** (Heck, 2015, Theorem 4.11)**.** $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\Theta)$ *proves* $\mathsf{Con}(\Theta)$. *Moreover, if* $\Theta$ *is finitely axiomatized, then* $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \Theta$ *proves* $\mathsf{Con}(\Theta)$.

So the disentangled framework still allows us to carry out the soundness argument.

---

[6] Leigh and Nicolai (2013) also develop this approach.

[7] An additional advantage of this way of proceeding is that the target theory could also be ZF, say, without that affecting the syntactic theory in any dramatic way. The syntax will talk about '$\epsilon$' instead of '$0$' and '$S$', but it will otherwise be unchanged.

[8] Halbach and Leigh (2024) discuss such theories at length.

[9] I borrow this notation from Leigh and Nicolai (2013).

[10] $\mathsf{Tr}(\Theta)$ is the formalization of "All of $\Theta$'s axioms are true". There will be many possible formalizations, in fact, depending upon how the set of $\Theta$'s axioms is represented. If $\Theta$ is finitely axiomatizable, then there is a canonical choice: $x = A_1 \vee \cdots \vee x = A_n$. But if it is not, then there famously is no canonical choice (Feferman, 1960). So problems of intensionality will arise here, as they do in the case of the second incompleteness theorem.

This result allows us to illustrate another advantage of 'disentangling'. The theory $\mathsf{CT}[\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA})]$ is inconsistent, because $\mathsf{CT}[\mathsf{I}\Sigma_1]$ proves $\mathsf{Con}(\mathsf{PA})$ already. So we cannot, in that way, formalize a (putative) soundness proof for $\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA})$. But we can do so in the disentangled framework: $\mathsf{Con}(\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA}))$ is provable in $\mathsf{CTD}[\mathsf{I}\Sigma_1] + (\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA}))$, which is consistent—since, by Theorem 2 below, it is mutually interpretable with $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA}))$. And how do we know that $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1 + \neg\mathsf{Con}(\mathsf{PA}))$ is consistent? Because it is true![11]

Does the disentangled framework solve the problems that motivated this framework in the first place? Heck (2015) does prove some limitative results that are meant to address this question. For our purposes, the most important of these is:

**Theorem 2** (Heck, 2015, Corollary 4.14)**.** $\mathsf{CTD}[\mathsf{I}\Sigma_n] + \mathsf{Tr}(\Theta)$ *is mutually interpretable with* $\mathsf{I}\Sigma_n + \mathsf{Con}(\Theta)$*. Moreover, if* $\Theta$ *is finitely axiomatized, then* $\mathsf{CTD}[\mathsf{I}\Sigma_n] + \Theta$ *is mutually interpretable with* $\mathsf{I}\Sigma_n + \mathsf{Con}(\Theta)$.

This implies, in particular, that $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_1$ is mutually interpretable with $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{I}\Sigma_1)$, which is a sub-theory not only of PA but of $\mathsf{I}\Sigma_2$. So if you want to prove $\mathsf{Con}(\mathsf{I}\Sigma_2)$, you do have to work in a stronger theory, namely $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{I}\Sigma_2$. The 'semantic' induction axioms in $\mathsf{CTD}[\mathsf{I}\Sigma_1]$ thus do not inadvertently strengthen the object-language theory.

However, this does not, by itself, answer all the questions one might want to ask here. One might wonder not just whether semantic induction stengthens the target theory but also whether our assumptions about the semantics of the target theory somehow infect the syntactic theory. Consider, for example, $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{PA})$. We can, of course, reason about the language of arithmetic in PA itself, via Gödel numbering, and prove facts about that language in PA that we cannot prove in $\mathsf{I}\Sigma_1$. But the semantic theory establishes a mapping from numerals to numbers: Each numeral $\overline{n}$ denotes the corresponding number $n$; so the open term "$\mathsf{den}(\overline{x})$"—the denotation of the $x^{\text{th}}$ numeral—describes a function that embeds the numbers in the syntactic sort, the 'snumbers', into the numbers in the target sort. Might this somehow allow us to transfer facts about the numbers in the target sort back into the syntax? If we can prove that all *numbers* have some (coded) syntactic property, and if the *snumbers* are provably mapped onto an initial segment of the numbers, might that imply that the *snumbers* have the (coded) syntactic property, too?

Theorem 2 does tell us that $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{PA})$ is mutually interpretable with $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{PA})$. But *that* theory does interpret PA, by the arithmetized

[11] Feferman (1960, Theorem 6.6) showed that PA interprets $\mathsf{PA} + \neg\mathsf{Con}(\mathsf{PA})$. His proof relies upon the reflexivity of PA, but Visser (2014) has proved that $\mathsf{I}\Sigma_n$ interprets $\mathsf{I}\Sigma_n + \neg\mathsf{Con}(\mathsf{I}\Sigma_n)$—and a more general result still. In light of remarks below, $\mathsf{I}\Sigma_n$ will therefore prove $\mathsf{Con}(\mathsf{I}\Sigma_n) \equiv \mathsf{Con}(\mathsf{I}\Sigma_n + \neg\mathsf{Con}(\mathsf{I}\Sigma_n))$.

completeness theorem,[12] so it isn't clear, at this point, whether we might be able to prove all of the *purely syntactic* induction axioms (as opposed to the ones with semantic vocabulary). The worry isn't, then, that $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$ is no different from $\mathsf{CTD}[\mathsf{PA}]+\mathsf{Tr}(\mathsf{PA})$. It seems unlikely, to say the least, that we'll be able to prove all the *extended* induction axioms, since our target theory has no semantic vocabulary. But the non-extended induction axioms—that is, the induction axioms in the purely syntactic language $\mathscr{S}$—are a different matter. The question is whether we can prove those. We *can*, after all, prove their analogue in the object-theory.

As we'll see, this worry can be addressed.

We'll start with a relatively simple case. Suppose we could prove all the induction axioms for $\mathscr{S}$ in $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$. Then, one might reason, since we can also prove $\mathsf{Con}(\mathsf{PA})$, and since $\mathsf{Con}(\mathsf{PA})$ is itself sentence of $\mathscr{S}$, $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$ would be a sub-theory of $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$. So it will be enough to show that $\mathsf{I}\Sigma_1+\mathsf{Con}(\mathsf{PA})$ does not interpret $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$. And this is easy to show.[13] Since PA is essentially reflexive, so is $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$, which means that $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$ proves the consistency of every one of its finite sub-theories. But $\mathsf{I}\Sigma_1+\mathsf{Con}(\mathsf{PA})$ is such a sub-theory, so $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$ proves $\mathsf{Con}(\mathsf{I}\Sigma_1+\mathsf{Con}(\mathsf{PA}))$. Since no (consistent) theory can interpret any theory that proves its consistency,[14] it thus follows that $\mathsf{I}\Sigma_1+\mathsf{Con}(\mathsf{PA})$ does not interpret $\mathsf{PA}+\mathsf{Con}(\mathsf{PA})$.

In fact, however, there is a subtlety here that we have overlooked. Volker Halbach once observed that we need to distinguish the consistency statement $\mathsf{Con}_{\mathscr{S}}(\Theta)$ in the *syntactic* language from the (coded) consistency statement $\mathsf{Con}_{\mathscr{A}}(\Theta)$ in the target language. What $\mathsf{CTD}[\Sigma]+\Theta$ proves is $\mathsf{Con}_{\mathscr{S}}(\Theta)$. Halbach additionally observed that $\mathsf{CTD}[\Sigma]+\Theta$ is always a conservative extension of $\Theta$, so $\mathsf{CTD}[\Sigma]+\Theta$ does *not* prove $\mathsf{Con}_{\mathscr{A}}(\Theta)$, unless $\mathsf{CTD}[\Sigma]+\Theta$ is inconsistent (Leigh and Nicolai, 2013, §3.2; Heck, 2018, §5). So we cannot transfer results from the syntax to the target.

Similarly, then, the statement $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA})$ that $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$ proves is a statement about the *arithmetical* version of PA. So it is really $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{A}})$ that $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$ proves, *not* $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$,which asserts the consistency of the *syntactic* version of PA. Now, on reflection, this does not affect the argument two paragraphs back: That depended only upon the fact that $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{A}})$ belongs to the language $\mathscr{S}$. But it does raise the question whether $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$ can also be proven. This might seem trivial and unimportant. But we'll need to know below that $\mathsf{CTD}[\mathsf{I}\Sigma_1]+\mathsf{Tr}(\mathsf{PA})$

[12] Indeed, $\mathsf{Q}+\mathsf{Con}(\mathcal{T})$ always interprets $\mathcal{T}$.

[13] Thanks to Carlo Nicolai for this observation.

[14] In fact, something stronger is true: Even if $\Theta$ only proves $\mathsf{Con}(\Sigma)$ on a cut, $\Sigma$ cannot interpret $\Theta$. See Heck (2015, §2.5) for a proof of this folklorish result, which is implicit in Pudlák (1985).

proves $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$, and the issue is more obviously important, and not at all trivial, if we take our syntactic theory to be, as it ideally should be, a theory of concatenation. Then $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$ does not even make sense, since PA is not a theory of concatenation. Similar issues arise when the object-language is, say, the langauge of set-theory. $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{ZF})$ is mutually interpretable with $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{ZF})$. It may not be immediately obvious whether this theory proves $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$.

However, $\mathsf{PA}^{\mathscr{S}}$ and $\mathsf{PA}^{\mathscr{A}}$ are (trivially) mutually interpretable. And $\mathsf{I}\Sigma_1$ is strong enough both (i) to show that $\mathsf{PA}^{\mathscr{S}}$ is interpretable in $\mathsf{PA}^{\mathscr{A}}$ and (ii) to prove that interpretability implies relative consistency, i.e., to prove that $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{A}})$ implies $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$. So $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{PA})$ does prove $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$, as wanted. This does not depend upon the triviality of the interpretation. Generally speaking, $\mathsf{I}\Sigma_1$ is strong enough to establish the existence of interpretations, where they exist, since this is just a matter of the existence of certain proofs.[15] So $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{ZF})$ does prove $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$, by showing that $\mathsf{PA}^{\mathcal{S}}$ is interpretable in ZF and so that $\mathsf{Con}(\mathsf{ZF})$ implies $\mathsf{Con}_{\mathscr{S}}(\mathsf{PA}^{\mathscr{S}})$. Similar results will be available when the syntax is the syntactic analogue of $\mathsf{I}\Sigma_1$:[16] Such a theory will be strong enough both to establish results about interpretability and to show that interpretability implies relative consistency. The argument that $\mathsf{I}\Sigma_1 + \mathsf{Con}(\mathsf{PA})$ does not interpret $\mathsf{PA} + \mathsf{Con}(\mathsf{PA})$ will thus generalize in the right way.

We now turn to the question whether even the $\Sigma_2$ induction axioms for $\mathscr{S}$ are provable in $\mathsf{CTD}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathsf{PA})$ (that is, whether $\mathsf{I}\Sigma_2^{\mathscr{S}}$ is a sub-theory thereof). We shall prove something stronger and more general:

**Theorem 3.** $\mathsf{CTD}[\mathsf{I}\Sigma_n^{\mathscr{S}}] + \mathsf{I}\Sigma_{n+1}^{\mathscr{A}}$ *does not contain* $\mathsf{I}\Sigma_{n+1}^{\mathscr{S}}$ *as a sub-theory (for* $n > 0$*).*

The beginning of the argument is the same as before. Suppose otherwise. Then, since $\mathsf{CTD}[\mathsf{I}\Sigma_n^{\mathscr{S}}] + \mathsf{I}\Sigma_{n+1}^{\mathscr{A}}$ proves $\mathsf{Con}_{\mathscr{S}}(\mathsf{I}\Sigma_{n+1}^{\mathscr{A}})$, it also proves $\mathsf{Con}_{\mathscr{S}}(\mathsf{I}\Sigma_{n+1}^{\mathscr{S}})$, so it contains $\mathsf{I}\Sigma_{n+1}^{\mathscr{S}} + \mathsf{Con}_{\mathscr{S}}(\mathsf{I}\Sigma_{n+1}^{\mathscr{S}})$ as a sub-theory. Since $\mathsf{CTD}[\mathsf{I}\Sigma_n^{\mathscr{S}}] + \mathsf{I}\Sigma_{n+1}^{\mathscr{A}}$ is mutually interpretable with $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$, it will suffice to establish:

[15] When the interpreted theory is not finitely axiomatizable, this becomes a bit more subtle, and may depend upon how the set of axioms is described, as Feferman (1960) famously shows. I'll ignore such subtleties here.

[16] It's most natural to take 'bounded' in a theory of concatentation to be defined in terms of substrings, as Halbach and Leigh (2024, §8.4) do. We can then define $\Sigma_n$, etc, in the usual way, and show that the syntactic theory with $\Sigma_n$ induction is mutually interpretable with $\mathsf{I}\Sigma_n$. This is because the most natural interpretation of arithmetic in syntax interprets numbers by strings like $|$, $||$, etc, and the usual coding of syntax in arithmetic has it that, if $a$ is a substring of $b$, then the code of $a$ is less than the code of $b$. (This is the much discussed 'monotonicity' of the usual codings (Heck, 2007; Halbach and Visser, 2014a,b; Grabmayr and Visser, 2021).)

**Lemma 4.** $\mathsf{I}\Sigma_{n+1} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ *is not interpretable in* $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$.

This is true despite the fact that $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ does interpret $\mathsf{I}\Sigma_{n+1}$ itself, by the arithmetized completeness theorem. So, in a sense, this is best possible.

For the proof, we need the following important fact:

**Theorem 5** (Hájek and Pudlák, 1993, p. 106, Theorem 4.33)**.** $\mathsf{I}\Sigma_n$ *proves that the theory whose (arithmetical) axioms are the true* $\Pi_{n+1}$ *sentences is consistent.*

Here, the true $\Pi_{n+1}$ sentences are characterized using a 'partial' truth-predicate for $\Pi_{n+1}$ sentences: a formula $\mathsf{Tr}_{\Pi_{n+1}}(x)$ such that, if $A$ is (no more complex than) $\Pi_{n+1}$, $\mathsf{I}\Sigma_n$ (in fact, $\mathsf{I}\Sigma_1$) proves $A \equiv \mathsf{Tr}_{\Pi_{n+1}}(\ulcorner A \urcorner)$.[17]

Theorem 5 is what allows us to show that $\mathsf{I}\Sigma_{n+1}$ proves $\mathsf{Con}(\mathsf{I}\Sigma_n)$, the reason being that $\mathsf{I}\Sigma_n$ has a finite $\Pi_{n+2}$ axiomatization.[18] And $\mathsf{I}\Sigma_{n+1}$ will prove that all of those axioms are true by the sort of trivial argument given above: Let $A$ be one of the axioms; then $A$ (since $\mathsf{I}\Sigma_n$ is a sub-theory of $\mathsf{I}\Sigma_{n+1}$); but also $A \equiv \mathsf{Tr}_{\Pi_{n+1}}(\ulcorner A \urcorner)$; so $\mathsf{Tr}_{\Pi_{n+1}}(\ulcorner A \urcorner)$. So $\mathsf{I}\Sigma_{n+1}$ proves that $\mathsf{I}\Sigma_n$ is a sub-theory of $\mathsf{Tr}(\Pi_{n+2})$. But even Q knows that a sub-theory of a consistent theory is consistent. So $\mathsf{I}\Sigma_{n+1}$ proves that $\mathsf{I}\Sigma_n$ is consistent if $\mathsf{Tr}(\Pi_{n+2})$ is and so proves that $\mathsf{I}\Sigma_n$ is consistent.

The proof of Lemma 4 just extends this argument.

*Proof of Lemma 4.* $\mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ is a $\Pi_1$ sentence, and $\mathsf{I}\Sigma_{n+1} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ proves that it is a true one, by the trivial argument. So $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ also has a $\Pi_{n+2}$ axiomatization, and $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ proves that all those axioms are true. So, by the same reasoning just rehearsed, $\mathsf{I}\Sigma_{n+1} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ proves $\mathsf{Con}(\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1}))$. So $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ is not interpretable in $\mathsf{I}\Sigma_{n+1} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$.[19] □

So that completes the proof of Theorem 3.

It's important to note that the theory mentioned in Theorem 5 is the one containing as axioms those $\Pi_{n+1}$ sentences that $\mathsf{I}\Sigma_n$ *thinks* are true, not necessarily the ones that actually are true—though, since $\mathsf{I}\Sigma_n$ is sound, the sentences it thinks are true really are true.[20] To put it more

[17] For the details, see Hájek and Pudlák (1993, pp. 50–61).

[18] This is because the usual, non-finite axiomatization is $\Pi_{n+2}$—just check the complexity of the induction axioms—and the finitely many axioms can be chosen from among the usual axioms (Hájek and Pudlák, 1993, p. 78, Theorem 2.52). (In my courses on these matters, I often assign the following exercise: If some infinitely axiomatized theory $\mathcal{T}$ has a finite axiomatization, then it has a finite axiomatization all of whose axioms are among those of $\mathcal{T}$.)

[19] This is really a special case of Hajék and Pudlák's Corollary 4.34(ii). But it is worth filling in some of the details they omit.

[20] Moreover, in the standard model, the extension of $\mathsf{Tr}_{\Pi_n}(x)$ is exactly the set of true $\Pi_n$ sentences.

formally, what Theorem 5 says is that $\mathsf{I}\Sigma_n$ proves: There is no sequence of formulas, ending with $0 = 1$, each of which is either a true $\Pi_{n+2}$ sentence or a logical axiom or else follows from earlier members of the sequence by one of the rules of inference. The proof of Lemma 4 depends upon the fact that $\mathsf{I}\Sigma_{n+1} + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ proves that $\mathsf{I}\Sigma_n + \mathsf{Con}(\mathsf{I}\Sigma_{n+1})$ is a sub-theory of the theory whose axioms are the true $\Pi_{n+2}$ sentences. It is irrelevant whether that claim is itself true.

This allows us to establish a generalization of Lemma 4 that is worth stating separately:

**Corollary 6.** *Let $A$ be a $\Pi_{n+2}$ sentence that is consistent with $\mathsf{I}\Sigma_n$. Then $\mathsf{I}\Sigma_{n+1} + A$ is not interpretable in $\mathsf{I}\Sigma_n + A$ .*

*Proof.* $\mathsf{I}\Sigma_n + A$ is $\Pi_{n+2}$ axiomatized, and $\mathsf{I}\Sigma_{n+1} + A$ proves that all its axioms are true. So $\mathsf{I}\Sigma_{n+1} + A$ proves $\mathsf{Con}(\mathsf{I}\Sigma_{n+1} + A)$. □

We can be confident, then, that 'disentangling' does what it is supposed to do: Not only do the semantic induction axioms not inadvertently strengthen the target theory, but assuming the truth of the induction axioms in the target theory does not 'transfer' back into the syntactic theory, either. That, however, seems about as much as we are likely to be able to say: $\mathsf{CTD}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_n$ is not a conservative extension of $\mathsf{I}\Sigma_n$, or even of $\mathsf{CTD}[\mathsf{I}\Sigma_n]$, not even for $\Pi_1$ sentences: The whole point is that $\mathsf{CTD}[\mathsf{I}\Sigma_n] + \mathsf{I}\Sigma_n$ proves the purely syntactic $\Pi_1$ sentence $\mathsf{Con}_{\mathscr{S}}(\mathsf{I}\Sigma_{n+1})$, which is not provable in $\mathsf{I}\Sigma_n$ itself (unless, of course, $\mathsf{I}\Sigma_n$ is inconsistent).[21]

## References


Feferman, S. (1960). 'Arithmetization of metamathematics in a general setting', *Fundamenta Mathematicae* 49: 35–92.

Grabmayr, B. and Visser, A. (2021). 'Self-reference upfront: A study of self-referential Gödel numberings', *The Review of Symbolic Logic*. Forthcoming.

Hájek, P. and Pudlák, P. (1993). *Metamathematics of First-order Arithmetic*. New York, Springer-Verlag.

Halbach, V. and Leigh, G. E. (2024). *The Road to Paradox: A Guide to Syntax, Truth, and Modality*. Cambridge, Cambridge University Press. Forthcoming.

Halbach, V. and Visser, A. (2014a). 'Self-reference in arithmetic I', *The Review of Symbolic Logic* 7: 671–91.

—— (2014b). 'Self-reference in arithmetic II', *The Review of Symbolic Logic* 7: 692–712.

Heck, R. K. (2007). 'Self-reference and the languages of arithmetic', *Philosophia Mathematica* 15: 1–29. Originally published under the name "Richard G. Heck, Jr".


[21]Thanks to Volker Halbach, Graham Leigh, and Carlo Nicolai for helping me sort all this out, and to Matteo Zicchetti for inspiring me to do so.

—— (2009). 'The strength of truth-theories'. Originally published under the name "Richard G. Heck, Jr".

—— (2015). 'Consistency and the theory of truth', *Review of Symbolic Logic* 8: 424–66. Originally published under the name "Richard G. Heck, Jr".

—— (2018). 'The logical strength of compositional principles', *Notre Dame Journal of Formal Logic* 59: 1–33. Originally published under the name "Richard G. Heck, Jr".

Leigh, G. E. and Nicolai, C. (2013). 'Axiomatic truth, syntax and metatheoretic reasoning', *The Review of Symbolic Logic* 6: 613–36.

Łełyk, M. Z. (2022). 'Model theory and proof theory of the global reflection principle', *The Journal of Symbolic Logic* 88: 738–79.

Pudlák, P. (1985). 'Cuts, consistency statements and interpretations', *Journal of Symbolic Logic* 50: 423–41.

Tarski, A. (1956). 'The concept of truth in formalized languages', in J. Corcoran (ed.), *Logic, Semantics, and Metamathematics*. Indianapolis, Hackett, 152–278.

Visser, A. (2014). 'The interpretability of inconsistency: Feferman's theorem and related results'. Preprint.