Some Remarks on ‘Logical’ Reflection

Richard Kimberly Heck

Department of Philosophy and Department of Linguistics, Brown University

Abstract

Cezary Cieśliński has proved a result shows that highlights ‘logical reflection’: The principle that every logically provable sentence is true. He suggests further that this result has a good deal of philosophical significance, specifically for the so-called ‘conservativeness argument’ against deflationism. This note discusses the question to what extent Cieśliński’s result generalizes, and just how strong ‘logical reflection’ is, and suggests that the answers to these questions call the philosophical (though not the technical) significance of Cieśliński’s result into doubt.

All the arithmetical axioms of PA are true. So are all the logical axioms. And the rules of inference preserve truth. So all theorems of PA are true.

Such an argument is known as a ‘soundness proof’ for PA. Its conclusion is the so-called global reflection priniciple for PA:[1]

$$\mathsf{GRfn}_{\mathsf{PA}} \qquad \forall x[\mathsf{Bew}_{\mathsf{PA}}(x) \rightarrow \mathsf{Tr}(x)]$$

Soundness proofs are of interest for a number of reasons, including the fact that the notion of truth seems to occur essentially in them. The proof is itself an induction on the length of PA-proofs, the ‘inductive predicate’ being: $\mathsf{Tr}(x)$. So the most natural way to formalize the argument would be in the theory known as CT[PA], which is PA plus a Tarski-style compositional theory of truth, where the (newly added) semantic vocabulary is allowed to occur in the induction axioms.[2]

Cezary Cieśliński, however, has suggested that the extended induction axioms are doing only a very particular sort of work. Let $\mathsf{GRfn}_{\emptyset}$

[1] Here $\mathsf{Bew}_{\mathcal{T}}(x)$ is some ‘standard’ provability predicate for $\mathcal{T}$.

[2] In fact, the proof can be formalized straightfowardly in $\mathsf{CT}[\mathsf{I}\Sigma_1]$ (Heck, 2015, Theorem 3.20).

be the global reflection principle for logic itself, which says that every *logically* provable sentence is true:

$$\mathsf{GRfn}_{\emptyset} \qquad \forall x[\mathsf{Bew}_{\emptyset}(x) \to \mathsf{Tr}(x)]$$

And let $\mathsf{CT}^{-}[\mathsf{PA}]$ be PA plus a Tarski-style compositional theory of truth, but *without* any extension of the induction axioms. Cieśliński (2010b) proves:

Theorem 1. $\mathsf{CT}^{-}[\mathsf{PA}] + \mathsf{GRfn}_{\emptyset}$ *proves* $\mathsf{GRfn}_{\mathsf{PA}}$.

So, the thought is, the strong truth-theory $\mathsf{CT}[\mathsf{PA}]$ is only needed to prove reflection for logic.

The result is, of course, not open to doubt. Here, though, I want to make some observations about it.

1. The result is essentially trivial for finitely axiomatized theories.

2. The result does not extend smoothly to infinitely axiomatized theories other than PA.

3. In the case of PA, the proof depends essentially upon the way $\mathsf{CT}^{-}[\mathsf{PA}]$ is formulated.

4. A natural generalization of the result can be stated and proved in the context of 'disentangled' truth-theories, and its limitations become clear there.

I'll take these points in order. As we'll see, they bear upon the philosophical significance of Cieśliński's result, which I have also discussed elsewhere (Heck, 2025).

1 Finitely Axiomatized Theories

Here is a generalization of Cieśliński's result for finitely axiomatized theories.

Theorem 2. *Let $\mathcal{T}$ be finitely axiomatized. Then* $\mathsf{CT}^{-}[\mathcal{T}] + \mathsf{GRfn}_{\emptyset}$ *proves* $\mathsf{GRfn}_{\mathcal{T}}$.

We assume here that $\mathcal{T}$ is strong enough to formalize the syntax needed for $\mathsf{CT}^{-}[\mathcal{T}]$. So, in the arithmetical case, it would be enough to assume that $\mathcal{T}$ contains Q.[3] But the result is not specific to arithmetical theories. It would apply, e.g., to set theories, so long as they interpret Q.

Proof. We work in $\mathsf{CT}^-[\mathcal{T}] + \mathsf{GRfn}_\emptyset$. Suppose $\mathcal{T}$ proves A. Then logic proves $\bigwedge \mathcal{T} \rightarrow A$,[4] where $\bigwedge \mathcal{T}$ is the conjunction of the finitely many axioms of $\mathcal{T}$. By $\mathsf{GRfn}_\emptyset$, $\mathsf{Tr}(\bigwedge \mathcal{T} \rightarrow A)$, and the clause of $\mathsf{CT}^-[\mathcal{T}]$ concerning the conditional then yields $\mathsf{Tr}(\bigwedge \mathcal{T}) \rightarrow \mathsf{Tr}(A)$. But $\mathcal{T}$ certainly proves $\bigwedge \mathcal{T}$, and $\mathsf{CT}^-[\mathcal{T}]$ will prove all T-sentences for the language of $\mathcal{T}$, i.e.: $\mathsf{Tr}(\bigwedge \mathcal{T}) \equiv \bigwedge \mathcal{T}$. So $\mathsf{Tr}(\bigwedge \mathcal{T})$ and therefore $\mathsf{Tr}(A)$, as wanted. $\square$

Note that all we really needed were the clause for the conditional and the T-sentence for $\bigwedge \mathcal{T}$.

We can do even better if we take our target to be the *local* reflection principle:

$\mathsf{Rfn}_\mathcal{T}$ $\qquad$ $\mathsf{Bew}_\mathcal{T}(A) \rightarrow A$

This is a schema: one instance for each sentence A. Let $\mathsf{DT}^-[\mathcal{T}]$ be the 'disquotational' truth-theory that extends $\mathcal{T}$ with just the instances of $\mathsf{Tr}(A) \equiv A$ for sentences of the language of $\mathcal{T}$, so without induction extended. Then we have:

Theorem 3. *Let $\mathcal{T}$ be finitely axiomatized. Then $\mathsf{DT}^-[\mathcal{T}] + \mathsf{Rfn}_\emptyset$ proves $\mathsf{Rfn}_\mathcal{T}$.*

Proof. Fix some sentence A and reason in $\mathsf{DT}^-[\mathcal{T}] + \mathsf{Rfn}_\emptyset$. Suppose that A is $\mathcal{T}$-provable. Then $\bigwedge \mathcal{T} \rightarrow A$ is logically provable, so $\mathsf{Tr}(\bigwedge \mathcal{T} \rightarrow A)$. But we also have the T-sentence for that sentence, so $\bigwedge \mathcal{T} \rightarrow A$. And of course we can prove the antecedent. $\square$

What all of this suggests to me is that, despite its seeming innocence, 'logical reflection' is a very strong principle indeed.[5] Indeed, we have:

Proposition 4. *If $\mathcal{T}$ is consistent, then $\mathsf{CT}^-[\mathcal{T}]$ does not prove $\mathsf{GRfn}_\emptyset$.*

[3] As we'll see below, there is some unclarity about what precisely $\mathsf{CT}^-[\mathcal{T}]$ is. If it is a theory of satisfaction, then we also need to assume that $\mathcal{T}$ is 'sequential'—that it can interpret a reasonable theory of sequences.

[4] Depending upon the formalization of logic we are using, we may need more than Q for this step. If the system has a rule of conditional proof, then it is easy to construct the proof of $\bigwedge \mathcal{T} \rightarrow A$. Hilbert-style axiomatic systems pose more of a challenge. But $\mathsf{I}\Delta_0 + \mathsf{exp}$ will suffice for any reasonable proof system, and I suspect that $\mathsf{I}\Delta_0 + \omega_1$ will suffice for the great majority.

[5] See Cieśliński's Corollary 2 for additional reason in favor of this claim.

Proof. Suppose that $\mathcal{T}$ proves $\mathsf{GRfn}_\emptyset$. Then some finite fragment $\mathcal{F}$ of $\mathcal{T}$ does so. By theorem 2, $\mathsf{CT}^-[\mathcal{F}]$ therefore proves $\mathsf{Con}(\mathcal{F})$. But that is impossible since $\mathsf{CT}^-[\mathcal{F}]$ is always a conservative extension of $\mathcal{F}$. □

Note that proposition 4 holds for *arbitrary* theories $\mathcal{T}$, even extremely powerful theories like ZFC plus whatever large cardinal axioms you would like to add. We don't even need to assume that $\mathcal{T}$ is recursively axiomatizable. The result holds even if $\mathcal{T}$ is true arithmetic![6] What's needed to prove $\mathsf{GRfn}_\emptyset$ is thus not any amount of logical strength but an extension of $\mathcal{T}$'s axioms to allow for induction on the length of proofs, where the 'inductive formula' contains semantic predicates. And, as we shall see below, the amount of induction needed for the proof of $\mathsf{GRfn}_\emptyset$ is as minimal as it could possibly be.

2 Infinitely Axiomatized Theories

The case of infinitely axiomatized theories is quite different. One's first thought might be to proceed as follows. If A is a theorem of some infinitely axiomatized theory $\mathcal{T}$, then, for some finite set $\mathcal{F}$ of axioms of $\mathcal{T}$, logic proves $\bigwedge \mathcal{F} \rightarrow A$. But…what? We can easily show that *each* axiom of $\mathcal{T}$ is true, by deriving it from the axiom itself and the T-sentence for it. We can even show that each finite collection of axioms is true. But we have no way to show that *all* axioms of $\mathcal{T}$ are true and therefore have no way to show that some unspecified finite collection of axioms is true.[7]

Cieśliński's proof shows that this problem can be side-stepped in the case of PA. Most of the action in the proof is in the demonstration that an *arbitrary* induction axiom is true, so I will simply the proof by focusing on that case.[8] Again, for *each* induction axiom, we can easily show that it is true. But here we are dealing with the induction axiom for some formula $A(x)$ or other, which has not been specified. So we cannot simply appeal to the T-sentence for this axiom, since we do not know which T-sentence that is. To put it differently: For all we know, $A(x)$ might be 'non-standard'; if it is, then then the T-sentence for the induction axiom for $A(x)$ isn't provable in $\mathsf{CT}^-[\mathsf{PA}]$, since there is no such axiom.

[6] It *is* being assumed that $\mathsf{CT}^-[\mathcal{T}]$ is 'typed', so that the semantic notions $\mathsf{CT}^-[\mathcal{T}]$ adds are not already present in the language of $\mathcal{T}$.

[7] Indeed, matters are even worse, as we shall see in section 4. Note that talk of finite sets can be formalized in PA through various coding mechanisms.

[8] Cieśliński's proof actually shows that arbitrary conjunctions of induction axioms are true, thus avoiding a problem to be mentioned in section 4.

Proposition 5. *Let* $\mathsf{Ind}(A)$ *be the induction axiom for the formula* A*:*

$$A(0) \wedge \forall x(A(x) \to A(Sx)) \to \forall x A(x)$$

Then $\mathsf{CT}^{-}[\mathsf{PA}] + \mathsf{GRfn}_{\emptyset}$ *proves* $\mathsf{Ind}(A)$.

Proof. We assume $\neg\mathsf{Tr}(\mathsf{Ind}(A))$, which by the clause for negation implies $\mathsf{Tr}(\neg\mathsf{Ind}(A))$, and attempt to reach a contradiction.

Clearly, logic proves:

$$\neg\mathsf{Ind}(A) \to [A(0) \wedge \forall x(A(x) \to A(Sx)) \wedge \exists x(\neg A(x))]$$

Since logic is true:

$$\mathsf{Tr}(\neg\mathsf{Ind}(A) \to [A(0) \wedge \forall x(A(x) \to A(Sx)) \wedge \exists x(\neg A(x))])$$

We then distribute the truth-predicate across the propositional connectives:

$$\mathsf{Tr}(\neg\mathsf{Ind}(A)) \to \mathsf{Tr}(A(0) \wedge \forall x(A(x) \to A(Sx))) \wedge \mathsf{Tr}(\exists x(\neg A(x)))$$

But we have assumed the antecedent, so:

$$\mathsf{Tr}(A(0) \wedge \forall x(A(x) \to A(Sx))) \wedge \mathsf{Tr}(\exists x(\neg A(x))) \tag{1}$$

At this point, Cieśliński (2010b, p. 413) writes that "by the properties of the truth-predicate" $\exists n(\mathsf{Tr}(\neg A(\overline{n}))$. We'll return to the question what this means.

Now, for any n, logic proves:

$$A(0) \wedge \forall x(A(x) \to A(Sx)) \to A(\overline{n})$$

That is: We can show, already in PA (and, indeed, in Q), that, for each n, we do not need to use induction to prove the displayed instance of $\mathsf{Ind}(A)$. We can do so simply by instantiating the second conjunct of the antecedent n times and using n applications of *modus ponens*. So, since logic is true:

$$\mathsf{Tr}(A(0) \wedge \forall x(A(x) \to A(Sx)) \to A(\overline{n}))$$

And distributing again:

$$\mathsf{Tr}(A(0) \wedge \forall x(A(x) \to A(Sx))) \to \mathsf{Tr}(A(\overline{n}))$$

But we had the antecedent at (1) above. So $\mathsf{Tr}(A(\overline{n}))$ and so $\forall n \mathsf{Tr}(A(\overline{n}))$, since n was arbitrary. But we also have $\exists n(\mathsf{Tr}(\neg A(\overline{n}))$ and, by the clause for negation, therefore $\exists n \neg\mathsf{Tr}(A(\overline{n}))$. So that's a contradiction. □

This proof is very specific to induction axioms.[9] Consider, for example, the theory $\mathsf{RCI(PA)}$ whose axioms are those of PA plus $\mathsf{Con(PA)}$, $\mathsf{Con}(\mathsf{PA} + \mathsf{Con(PA)})$, and so forth. I do not have a proof that one cannot prove in $\mathsf{CT}^-[\mathsf{RCI(PA)}] + \mathsf{GRfn}_\emptyset$ that all axioms of $\mathsf{RCI(PA)}$ are true, but I do not see how to prove that they are.[10] Cieśliński's proof relies essentially upon the fact that each instance

$$A(0) \wedge \forall x(A(x) \to A(Sx)) \to A(\overline{n})$$

of the induction axiom for $A(x)$ can be proven in pure logic. Nothing of the sort is true for the extra axioms of $\mathsf{RCI(PA)}$.

The point is clearer when we consider theories formulated in languages other than the language of arithmetic. Consider, for example, ZFC. I cannot see how even to begin to prove in $\mathsf{CT}^-[\mathsf{ZFC}] + \mathsf{GRfn}_\emptyset$ that *all* of the separation and replacement axioms are true (though we can easily enough prove that *each* of them is true). Nothing like Cieśliński's proof strategy is available here.

So PA looks like a very special case.

3 Truth and Satisfaction

And, even in that special case, the proof depends essentially upon how $\mathsf{CT}^-[\mathsf{PA}]$ is formulated. As I mentioned above, one of the key moves in the proof is from $\mathsf{Tr}(\exists n(\neg A(n)))$ to $\exists n(\mathsf{Tr}(\neg A(\overline{n}))$, which is meant to follow "by the properties of the truth-predicate". Cieśliński does not tell us, in this paper,[11] exactly what $\mathsf{CT}^-[\mathsf{PA}]$ is, but, if this move is to be legitimate,

[9] Cieśliński also does not consider induction axioms with parameters, and the proof would not work for that case. However, if we have full induction, then parameter-free induction is equivalent to induction with parameters (and that is provable in PA itself). But that is not the case when considering fragments: Σ_n parameter-free induction is weaker than Σ_n induction with parameters (Kaye et al., 1988). Thanks to Albert Visser for consulation and the reference.

[10] Below, I will show that $\mathsf{CT}^-[\mathsf{PA}] + \mathsf{GRfn}_\emptyset$ proves the 'formalized ω rule' for PA. If we had the same result here, then perhaps that would do it. But the proof of that result *depends upon* the fact $\mathsf{CT}^-[\mathsf{PA}] + \mathsf{GRfn}_\emptyset$ proves that all axioms of PA are true and, in fact, that all conjunctions of axioms of PA are true.

[11] Cieśliński (2010b, p. 412) denotes this theory $PA^-(S)$ and says that it "extend[s] the language of arithmetic with a new predicate 'Tr'" and "add[s] the usual Tarski clauses as new axioms". So he certainly has the 'pure' truth theory in mind. But nothing like ($\exists$Eq) was part of Tarski's theory, which involved satisfaction. In any event, Cieśliński (2010a, p. 326) uses the same notation in another paper, and there $PA^-(S)$ is the theory discussed in the text.

it must be a 'pure truth-theory', so to speak, rather than a theory of satisfaction. That is, the clause for the existential quantifier must be:

$$\mathsf{Tr}(\exists n A(n)) \equiv \exists n \mathsf{Tr}(A(\overline{n})) \tag{$\exists$Eq}$$

so that truth for quantified formulas is defined in terms of the truth of their instances. And if that is what Cieśliński is assuming, then there is no problem.

But, while it is technically very convenient to handle quantifiers this way when the theory in question is formulated in the language of arithmetic, we must remember that it is *just* a convenience: a technical simplification. When our interest is in philosophical questions, especially epistemological questions, it can obscure important facts. A statement of the form "Every natural number is F" does not *mean* that every instance $F(\overline{n})$ is true. The determiner "Every" has a meaning of its own, as does the predicate "natural number", and what "Every natural number" means is determined by the meanings of those two expressions, just as the meanings of "Every real number" and "Every horse" are so determined. It is, to be sure, *true* that "Every natural number is F" is true iff $F(\overline{n})$ is true, for all n, but that is a consequence of two more basic facts: what "Every natural number is F" means, which should be stated in terms of satisfaction, and the fact that every natural number has a numeral that names it.

The lesson is: ($\exists$Eq) is a *theorem*, epistemologically speaking, not something that should simply be built into the theory of truth. And the proof of ($\exists$Eq) is not trivial, even once we have proven that every number is denoted by a numeral.[12] We also need:

Lemma 6. *$F(x)$ is true when x is assigned the value n iff $F(\overline{n})$ is true. More generally, $F(\overline{n}, \vec{y})$ is satisfied by α iff $F(x, \vec{y})$ is satisfied by the sequence that is just like α but assigns n to x.*

This 'extensionality lemma' is not difficult to prove, but the proof is by induction on the complexity of F, and the induction obviously involves semantic machinery. The extensionality lemma therefore cannot be proven in $\mathsf{CT}^-[\mathsf{PA}]$, if that theory is formulated as a theory of satisfaction.

Indeed, in the presence of $\mathsf{GRfn}_\emptyset$, formulating $\mathsf{CT}^-[\mathcal{T}]$ as Cieśliński does amounts to building the 'formalized ω rule'

$$\forall n \mathsf{Bew}_\mathcal{T}(A(\overline{n})) \to \forall n A(n)$$

[12] Denotation for terms can be defined in PA itself, and it can be proven in $\mathsf{I}\Sigma_1$ that every number is denoted by a term.

into the semantics. Consider, e.g., $\mathsf{CT}^-[\mathsf{I}\Sigma_1]$ and, working therein, suppose that (the finitely axiomatizable theory) $\mathsf{I}\Sigma_1$ proves $\forall n \mathsf{Bew}_{\mathsf{I}\Sigma_1}(A(\overline{n}))$. Then $\forall n[\mathsf{Bew}_\emptyset(\bigwedge \mathsf{I}\Sigma_1 \rightarrow A(\overline{n}))]$, so by logical reflection $\forall n \mathsf{Tr}(\bigwedge \mathsf{I}\Sigma_1 \rightarrow A(\overline{n}))$. Distributing, and noting that $\bigwedge \mathsf{I}\Sigma_1$ is a sentence, we have: $\mathsf{Tr}(\bigwedge \mathsf{I}\Sigma_1)) \rightarrow \forall n \mathsf{Tr}(A(\overline{n})$. But of course the antecedent is provable, so $\forall n \mathsf{Tr}(A(\overline{n})$. So $\mathsf{Tr}(\forall n A(n))$, by the clause for the universal quantifier, and hence $\forall n A(n)$. A similar proof works for PA since Cieśliński shows us how to prove that arbitrary conjunctions of PA's axioms are true. But if we have the formalized ω rule, then it's no surprise that we can prove global reflection.

The same sort of point applies more obviously to the language of set theory. While one can formulate a 'pure truth-theory' for that language by extending the language with a constant for each set (Fujimoto, 2012), that leads to collections of terms and formulas that do not form sets, which means that even the *logical* axioms do not form a set. So that approach is extremely artificial. The natural theory of truth for the language of set theory is the one that uses satisfaction.[13] Note that this has nothing to do with the *strength* of the theory. It's as true for MacLane set theory (Zermelo set theory plus choice, with separation restricted to Δ_0 formulas) as it is for ZFC.

4 A General Statement of Cieśliński's Result

One might have thought that a proof of logical reflection should be independent of the language: Whether logical reflection holds does not depend upon whether we are talking about numbers or sets or graphs. In the usual setting, however, there is no way to achieve such independence, since the truth-theory is simply grafted onto the original language, which is what provides for the coding of syntax. In a 'disentangled' setting, however, we can achieve such independence, or at least get very close to it.

The basic idea behind 'disentangling' is to separate the meta-language, in which syntatic and semantic arguments are carried out, from the object-language, which is the subject of those arguments. So we are working in a many-sorted context, with one sort for the syntactic objects—

[13] Dean (2015, p. 56, fn. 37) suggests that there are obstacles even to formulating a theory of truth for the language of set theory, ones we can perhaps overcome as Fujimoto (2012) does. But there are no such obstacles. Tarski's original example, after all, is what he calls the 'calculus of classes'.

terms and formulae—another for assignments, and a third for whatever the object-theory concerns: numbers, sets, whatever (Heck, 2009; Leigh and Nicolai, 2013; Heck, 2015). The meta-language might then contain only a primitive expression for concatenation, along with primitive terms denoting the various primitive symbols of the object-language. It is, however, convenient to allow the meta-language to be an arithmetical language, so that syntax is done through coding, though this language is still meant to be distinct from the object-language, even when that is also an arithmetical language. But the key point is that the meta-language can be arithmetical even when the object-language is not—the latter might be the language of set theory. The meta-*theory* that provides the resources for carrying out syntactic and semantic arguments can then also be separated from the object-*theory* whose consistency, say, we are trying to prove.

So let $\mathsf{CTD}_{\mathcal{L}}[\mathcal{S}]$ be the disentangled theory of truth for the language $\mathcal{L}$ built on the meta-theory $\mathcal{S}$. So $\mathsf{CTD}_{\mathcal{L}}[\mathcal{S}]$ will contain the usual clauses for the logical connectives, a weak theory of variable assignments, and 'disquotational' axioms for the non-logical expressions in $\mathcal{L}$. For example, if $\mathcal{L}$ is the language of set theory, then $\mathsf{CTD}_{\mathcal{L}}[\mathcal{S}]$ will contain:

$$(\epsilon) \qquad \mathsf{Sat}_{\alpha}(x \in y) \equiv \mathsf{Val}_{\alpha}(x) \in \mathsf{Val}_{\alpha}(y)$$

If $\mathcal{S}$ contains axiom schemata (e.g., induction axioms), then $\mathsf{CTD}_{\mathcal{L}}[\mathcal{S}]$ extends those by allowing semantic vocabulary into them; by contrast, $\mathsf{CTD}^{-}_{\mathcal{L}}[\mathcal{S}]$ does not extend those schemata.

This framework allows us to state some generalizations and even strengthenings of Cieśliński's result. Here is a first.

Theorem 7. *If* $\mathcal{T}$ is a finitely axiomatized theory in the language $\mathcal{L}$, then $\mathsf{CTD}^{-}_{\mathcal{L}}[\mathsf{Q}] + \mathsf{GRfn}_{\emptyset} + \mathcal{T}$ *proves* $\mathsf{GRfn}_{\mathcal{T}}$.

What this result allows us to see is that the truth-theory needed is even weaker than in the earlier generalization of Cieśliński's version, at theorem 2 (where we had $\mathsf{CT}^{-}[\mathcal{T}]$). The underlying syntax is given by Q and so is about as weak as it could be. Weak though it is, however, $\mathsf{CTD}^{-}_{\mathcal{L}}[\mathsf{Q}]$ still proves all T-sentences for $\mathcal{L}$ (Heck, 2015, Lemma 4.2). Since $\mathcal{T}$ proves $\bigwedge \mathcal{T}$, that then allows us to prove $\mathsf{Tr}(\bigwedge \mathcal{T})$, which is the key to proving $\mathsf{GRfn}_{\mathcal{T}}$.

When $\mathcal{T}$ is not finitely axiomatized, simply assuming $\mathcal{T}$ as object-theory will not, in general, allow us to prove that *all* axioms of $\mathcal{T}$ are true, even though it will allow us to prove that each of them is. Nonetheless, one might think we could instead prove the following (compare Heck, 2015, theorem 4.11).

Wanted Theorem. $\mathsf{CTD}^-_{\mathcal{L}}[\mathsf{Q}] + \mathsf{GRfn}_{\emptyset} + \mathsf{Tr}(\mathcal{T})$ proves $\mathsf{GRfn}_{\mathcal{T}}$.

Here $\mathsf{Tr}(\mathcal{T})$ is the obvious formalization of: All of $\mathcal{T}$'s axioms are true. Note that the mentioned theory will contain $\mathcal{T}$, since each axiom of $\mathcal{T}$ can be proven to be an axiom of $\mathcal{T}$[14] and then derived from $\mathsf{Tr}(\mathcal{T})$ and the T-sentence for that axiom.

Purported Proof. Reason in the mentioned theory. Suppose $\mathcal{T}$ proves A. Then some finite fragment $\mathcal{F}$ of $\mathcal{T}$ proves A. So logic proves: $\bigwedge \mathcal{F} \rightarrow A$. So by $\mathsf{GRfn}_{\emptyset}$, $\mathsf{Tr}(\bigwedge \mathcal{F} \rightarrow A)$. By the clause for $\rightarrow$, $\mathsf{Tr}(\bigwedge \mathcal{F}) \rightarrow \mathsf{Tr}(A)$. But $\mathsf{Tr}(\mathcal{T})$ implies $\mathsf{Tr}(\bigwedge \mathcal{F})$, so $\mathsf{Tr}(A)$. $\square$

The problem with this proof comes at the end: Exactly how does $\mathsf{Tr}(\mathcal{T})$ imply $\mathsf{Tr}(\bigwedge \mathcal{F})$? Of course $\mathsf{Tr}(\mathcal{T})$ implies $\mathsf{Tr}(\mathcal{F})$: All the axioms in $\mathcal{F}$ are true. But that is not the same as saying that the *conjunction* of the sentences in $\mathcal{F}$ is true. Given any specific finite set of sentences, it is easy enough to show that, if they are all true, then their conjunction is true. But $\mathcal{F}$ is not a specific set of sentences but some arbitrary finite set, so we have no way of simply conjoining its members. What we seem to need is something like:

(CC) If every element of the finite set $\mathcal{F}$ is true, then the conjunction of $\mathcal{F}$'s members is true.

The obvious proof of (CC) would be by induction on the size of the set (or the length of the conjunction), and such an induction is not even available in $\mathsf{CTD}^-_{\mathcal{L}}[\mathsf{PA}]$, since the induction axioms have not been extended. What we would need for the proof is $\mathsf{CTD}_{\mathcal{L}}[\mathsf{I}\Delta_0]$. Moreover, the principle of 'conjunctive correctness'—if every sentence in the finite set X is true, then the conjunction of members of X is true—is surprisingly strong (Enayat and Pakhomov, 2019): $\mathsf{CT}^-[\mathsf{I}\Delta_0 + \mathsf{exp}]$+(CC) proves $\mathsf{GRfn}_{\mathsf{PA}}$, so (CC) is equivalent to $\mathsf{GRfn}_{\mathsf{PA}}$ over $\mathsf{CT}^-[\mathsf{I}\Delta_0 + \mathsf{exp}]$. While I do not have a proof, then, that the Wanted Theorem cannot be proven, there is some reason to doubt that it can be.

[14] We will need to requite that the formula that defines the axioms of $\mathcal{T}$ 'represents' that set in the usual technical sense. Note that there will, in fact, be many way of representing that set and so many such formalizations. In the finite case, there is a canonical choice: $x = A_1 \vee \cdots \vee x = A_n$. But in the infinite case, there is famously no canonical choice (Feferman, 1960). So problems of intensionality will arise here, as they do in the case of the second incompleteness theorem.

If not, however, then the best we can get by way of generalizing Cieśliński's result would seem to be something like:

Theorem 8. $\mathsf{CTD}_{\mathcal{L}}[\mathsf{I}\Delta_0] + \mathsf{GRfn}_{\emptyset} + \mathsf{Tr}(\mathcal{T})$ *proves* $\mathsf{GRfn}_{\mathcal{T}}$.

But $\mathsf{GRfn}_{\emptyset}$ is just redundant here, since we already have:[15]

Theorem 9. $\mathsf{CTD}_{\mathcal{L}}[\mathsf{I}\Delta_0] + \mathsf{Tr}(\mathcal{T})$ *proves* $\mathsf{GRfn}_{\mathcal{T}}$.

Moreover, if we take $\mathcal{T}$ to be empty, then we have:

Theorem 10. $\mathsf{CTD}_{\mathcal{L}}[\mathsf{I}\Delta_0]$ *proves* $\mathsf{GRfn}_{\emptyset}$.

Which answers the question what is needed to prove $\mathsf{GRfn}_{\emptyset}$.[16] But we can improve that result. The only semantic clauses needed for the proof are those for the *logical* part of the language. For example, if $\mathcal{L}$ is the language of set theory, then we do not need the clause (ϵ) mentioned above. The same will be true whenever $\mathcal{L}$ is relational, i.e., contains no terms other than variables. If $\mathcal{L}$ does contain such terms, then we will need to know that every term has a denotation (under every assignment) in order to justify Universal Instantiation and Existential Generalization; but that is all we need to know for the proof of theorem 10. So, when there are no terms that are not variables, we have a result we might state as:

Theorem 11. $\mathsf{CTD}_{\emptyset}[\mathsf{I}\Delta_0]$ *proves* $\mathsf{GRfn}_{\emptyset}$.

And we will have a corresponding result for languages with non-variable terms, though we shall have to add the mentioned assumption that every term denotes.

5 Concluding Remarks

As technically significant, and elegant, as Cieśliński's result is, it does not generalize. It does apply to finitely axiomatized theories, but that hardly needed proving. But it does not seem to apply, in any sensible way, to infinitely axiomatized theories generally. Not only is there no

[15] A version of theorem 9 is implicit in Heck (2015, Theorem 4.11). What is proven there is that $\mathsf{CTD}_{\mathcal{L}}[\mathsf{I}\Sigma_1] + \mathsf{Tr}(\mathcal{T})$ proves $\mathsf{Con}(\mathcal{T})$, but the proof is just a soundness argument, so it actually establishes $\mathsf{GRfn}_{\mathcal{T}}$. Łełyk (2022) shows that $\mathsf{I}\Sigma_1$ can be replaced by $\mathsf{I}\Delta_0$.

[16] See Heck (2025, §2.2.2) for discussion of whether 'logical reflection' might be justified in some other way.

reason to believe that we can prove, say, that all axioms of ZFC are true in $\mathsf{CT}^{-}[\mathsf{ZFC}] + \mathsf{GRfn}_{\emptyset}$, even if we *assume* that all axioms of ZFC are true, and so work in $\mathsf{CT}^{-}[\mathsf{ZFC}] + \mathsf{GRfn}_{\emptyset} + \mathsf{Tr}(\mathsf{ZFC})$, there is still a significant obstacle to proving $\mathsf{GRfn}_{\mathsf{ZFC}}$, namely, something that will do the work of (CC). Perhaps these gaps can be filled. If not, however, then Cieśliński's result would seem to be of limited philosophical significance: It works for PA, but PA is a very special case.[17]

References

Cieśliński, C. (2010a). 'Deflationary truth and pathologies', *Journal of Philosophical Logic* 39: 325–37.

—— (2010b). 'Truth, conservativeness, and provability', *Mind* 119: 409–42.

Dean, W. (2015). 'Arithmetical reflection and the provability of soundness', *Philosophia Mathematica* 23: 31–64.

Enayat, A. and Pakhomov, F. (2019). 'Truth, disjunction, and induction', *Archive for Mathematical Logic* 58: 753–766.

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

Fujimoto, K. (2012). 'Classes and truths in set theory', *Annals of Pure and Applied Logic* 163: 1484–523.

Heck, R. K. (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".

—— (2025). 'Truth, soundness, and implicit commitment'. Draft.

Kaye, R., Paris, J., and Dimitracopoulos, C. (1988). 'On parameter free induction schemas', *Journal of Symbolic Logic* 53: 1082–97.

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

[17] Thanks to Cezary Cieśliński and Albert Visser for discussion.

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