MODELS OF POSITIVE TRUTH

MATEUSZ ŁEŁYK
Institute of Philosophy, University of Warsaw
and
BARTOSZ WCISŁO
Institute of Mathematics, University of Warsaw

Abstract. This paper is a follow-up to [4], in which a mistake in [6] (which spread also to [9]) was corrected. We give a strengthening of the main result on the semantical nonconservativity of the theory of PT⁻ with internal induction for total formulae (PT⁻ + INT(tot), denoted by PT⁻ in [9]). We show that if to PT⁻ the axiom of internal induction for all arithmetical formulae is added (giving PT⁻ + INT), then this theory is semantically stronger than PT⁻ + INT(tot). In particular the latter is not relatively truth definable (in the sense of [11]) in the former. Last but not least, we provide an axiomatic theory of truth which meets the requirements put forward by Fischer and Horsten in [9]. The truth theory we define is based on Weak Kleene Logic instead of the Strong one.

§1. Introduction.

1.1. Axiomatic theories of truth. Axiomatic theories of truth\(^1\) is a branch of mathematical logic and philosophy which studies the properties of formal theories generated in the following way:

1. We take a base theory \(B\) which we demand to be sufficiently strong to (strongly) represent basic syntactical operations.

2. We extend the language of \(B\) by adding one new unary predicate \(T\), and some axioms for it so that the resulting theory \(\text{Th}\) proves all sentences of the form

\[
T(⌜\phi⌝) \equiv \phi
\]

for \(\phi\) in the language of our base theory \(B\).

For the purpose of investigating this question we focus on the truth theories with Peano Arithmetic as a base theory. Various axiomatizations for the newly added predicate \(T\) correspond to the properties the (natural-language) notion of truth has. For example it can be compositional with respect to some set of connectives or untyped. By investigating various metamathematical properties of so constructed theories we hope to have a better grasp of how various properties of the notion of truth contribute to its meaning. In the current study we give support to the research initiated in [9] of justifying that the notion of truth has an expressive role. Moreover, we investigate the theory (together with its variant)

---

\(^1\) For a brief introduction to the subject see [14] and for a more complete one—[12].
which played the key role in [9] (and was earlier studied in [6] and [7]). However, we feel that our motivations are best summarized as the efforts to answer the following big question:

How various axioms for the truth predicate influence its strength?

1.2. The notion of strength. The notion of strength may enjoy many different explanations. For example, the simplest one is given by inclusion of sets of consequences: we might say that Th₁ is not weaker than Th₂ if and only if Th₁ proves all the axioms of Th₂. For many applications this is too fine-grained: many pairs of theories, such that one is intuitively stronger than the other, become incomparable.² A better adjusted notion was introduced by Kentaro Fujimoto in [11] and is a special kind of interpretability. We recall the definition:

**Definition 1.1.** Let Th₁ and Th₂ be axiomatic truth theories and let T_{Th₂} be the truth predicate of Th₂.

1. For any sentence Θ of L_{Th₂} and a formula φ(x) ∈ L_{Th₁} with precisely one free variable let

   Θ[φ(x)/T_{Th₂}(x)]

   denote the L_{Th₁} sentence which results from Θ by substituting φ(t) for every occurrence of T_{Th₂}(t), for an arbitrary term t, possibly with free variables. We rename the bounded variables if necessary.

2. We say that Th₁ relatively truth defines Th₂ if and only if there exists a formula φ(x) ∈ L_{Th₁} such that for any axiom Θ of Th₂³

   Th₁ ⊢ Θ[φ(x)/T_{Th₂}(x)].

   If Th₂ relatively truth defines Th₁ we will denote it by Th₁ ≤₆ Th₂.⁴

In terms of interpretations relative truth definability is an L_{PA}-conservative interpretation between truth theories (for the terminology related to interpretations see, e.g., [6]). Such an interpretation will be called also ω-interpretation. It was argued in [11] that relative truth definability provides a reasonably good explication of conceptual reduction between truth theories.⁵ We may treat it as an explication of the notion of strength: Th₁ is Fujimot-stronger than Th₂ if and only if Th₁ relatively truth defines Th₂ but not vice versa. This relation will be denoted by ≤₆.

1.2.1. Strength relative to PA. In some philosophical debates, especially the ones related to the deflationism, the need for a differently oriented formal explication of strength seems to emerge. It has been claimed (most importantly in [16, 19, 24]) that deflationary thesis that truth is a “simple” (aka “light”, “metaphysically thin”) notion implies that the

---

² Good examples are TB and CT⁻ (defined as CT↾ in [12]). The first one admits very weak axioms for the truth predicate, but is fully inductive (the induction scheme is extended to the language with the T predicate). The second admits compositional axioms for the truth predicate, but the induction scheme is restricted to arithmetical formulae only.

³ As pointed out by the anonymous referee, this definition is only suitable for theories formulated in logic closed under Modus Ponens. Since all theories which we consider are formulated in the classical logic, this will not be a real restriction for us.

⁴ “F” stands for “Fujimoto.”

⁵ Note that in Fujimoto’s own words this notion might still be ‘too coarse’ to provide a fully satisfactory explication of conceptual reducibility.
deflationary theory of truth should be conservative over PA. Let us recall that a theory of truth can be conservative over PA in two senses:

**Definition 1.2.** Let Th be a theory of truth.

1. We say that Th is proof-theoretically conservative over PA if and only if for every φ ∈ LPA, if Th ⊢ φ, then PA ⊢ φ.\(^7\)
2. We say that Th is model-theoretically conservative over PA if and only if every model M of PA admits an expansion to a model of Th.\(^8\)

**Remark 1.3.** Note that in the definition of model-theoretical conservativity we do not merely demand every model to have an extension to a model of Th, in which case both notions of conservativity would be the same. We say that M’ is an expansion of M if M and M’ are the same model, except that M’ carries interpretation of additional function, relation or constant symbols.

The two notions lead in the natural way to the following generalizations:

**Definition 1.4.** Let Th₁ and Th₂ be two truth theories.

1. We say that Th₁ is proof-theoretically not stronger than Th₂ if every LPA sentence provable in Th₁ is provable in Th₂. If Th₁ is proof-theoretically not stronger than Th₂, we will denote it with Th₁ \(\leq_P\) Th₂.\(^9\)
2. We say that Th₁ is model-theoretically not stronger than Th₂ if every model which can be expanded to a model of Th₂, can be expanded to a model of Th₁ as well. If Th₁ is model-theoretically not stronger than Th₂, we will denote it with Th₁ \(\leq_M\) Th₂.\(^10\)

Obviously, we say that Th₂ is proof-theoretically (model-theoretically) stronger than Th₁ if Th₁ \(\leq_P\) Th₂ but Th₂ \(\not\leq_P\) Th₁ (respectively, Th₁ \(\leq_M\) Th₂ but Th₂ \(\not\leq_M\) Th₁). This relation will be denoted \(\leq_P\) (\(\leq_M\) respectively).

Let us observe that the three notions of strength introduced above can be ordered with respect to their “granularity”. Indeed, for any theories Th₁ and Th₂ we have:

\[
Th₁ \leq_F Th₂ \implies Th₁ \leq_M Th₂ \implies Th₁ \leq_P Th₂. \quad \text{(FMP)}
\]

Hence, also

\[
Th₂ \not\leq_P Th₁ \implies Th₂ \not\leq_M Th₁ \implies Th₂ \not\leq_F Th₁. \quad \text{(-PMF)}
\]

Having three different notions of strength makes it possible to decide not only whether one theory of truth is stronger than another one, but also how much stronger it is.

Although the philosophical interest in the above notions of conservativity originated from investigations into deflationism we think that their philosophical importance and applicability are not limited to the debate over this stance. One of the reasons is their connection to the notion of relative truth definability, which was already stressed and which is of independent interest. Furthermore, the notion of model-theoretical strength can be

---

\(^6\) This thesis, however, has been recently criticised at length in [2].

\(^7\) This property is also called syntactical conservativity.

\(^8\) This relation is also called semantical conservativity.

\(^9\) \(P\) is meant to abbreviate “Proof”.

\(^10\) \(M\) is meant to abbreviate “Model”.

used in the study of formal explication of the notion of expressivity, as was used in [9] (in which the deflationary stance plays only an auxiliary role).

The third example of implementation is the comparison between classical and nonclassical logics and between various nonclassical logics. For example, two natural axiomatizations of Kripke fixed point construction (from [20]), i.e., KF (introduced in [5]) and PKF (introduced in [13]) differ dramatically in proof-theoretic strength (as discovered in [13]). In [15] it was claimed that the only reason for this is the use of classical logic as the external logic in KF. This thesis was justified by calibrating the proof-theoretical strength of the subsystems of KF and PKF in which the induction is restricted to arithmetical signature. In such a way the research on the strength of axiomatic theories shed some light on the use of classical logic in mathematics. Last but not least, the strength of axiomatic theories of truth served as a way to compare between Weak and Strong Kleene Logic (we say more about them in the next subsection). This was motivated by the question of Volker Halbach (see [12]) on whether we can justify calling the two logics “weak” and “strong” by looking at strength relations between axiomatic theories of truth?11 One of the contributions of the current paper consists in pointing at two theories of truth, which differ only in that one is based on Weak and the second one on Strong Kleene Logic, and proving that the former version is indeed semantically weaker than the latter. This implies that the Strong Kleene version is not relatively truth definable in its Weak Kleene counterpart.

All in all, we believe that the very question:

Which principles make the notion of truth strong?

is of independent philosophical interest. It is worth studying in order to better understand both the concept of “truth” and the concept of “strength”, which, in turn, are central to a number of discussions in formal philosophy.

Finally, the abovementioned reducibility notions (proof- and model-theoretical conservativity and relative truth definability) are very natural from the purely logical perspective, hence providing separating examples for them can (and moreover: should) be seen as an utterly independent research project.

1.3. The expressive power of truth. In [9] the authors searched for a theory of truth that would simultaneously satisfy two requirements:

1. It could model the use of truth in model theory;
2. It would witness the expressive function of the notion of truth.

The way to satisfy the former is to be model-theoretically conservative over the background theory.12 Being such, the theory would not discriminate among possible interpretations of our basis theory. In the authors’ own words:

But if we take the general model theoretic (or algebraic) perspective and treat all models as being on a par, then a truth theory for a background theory should not eliminate possibilities, for then it would not be a truth theory for the background theory (conceived of as a class of models). ([9], p. 352)

11 More precisely, the question of Halbach considered relative truth definability of KF in WKF.
12 As in our case, the authors focus on theories of truth extending PA.
Hence, the need for a model-theoretically conservative theory of truth flows from the “algebraic” view of theories: a theory is identified with the set of its models. The authors give no argument supporting such an approach, but simply observe that this is a viewpoint that is commonly adopted in certain parts of mathematics such as model theory. ([9], p. 364)

Concerning the second requirement authors seek for a source of expressivity which would not imply that the respective theory is a nonconservative extension of its background theory. They list the following metamathematical properties of a theory which witness that the respective notion adds expressiveness to the basic language:13

1. not being relatively interpretable in its background theory;
2. being a finitely axiomatizable extension of a theory which is not finitely axiomatizable;
3. having the superexponential speed-up over its background theory;
4. formalizing important semantical metatheorems (such as the completeness theorem).

It is not the purpose of this paper to discuss the philosophical relevance of the above criteria: for a careful discussion we refer the reader to the original paper. What we aim at is to provide a theory of truth which would fill in the gap caused by the discovery of the model-theoretic nonconservativity of PT₁⁺ INT↾tot. This is done in §4.

1.4. Compositional positive truth and its extensions. Before continuing let us introduce some handy notational conventions:

Convention 1.5. 1. For definiteness we assume that PA is formulated in the language with + and × as the only primitive symbols.

13 The list below is taken from the conclusion section of [9]. Earlier in the paper a more general criterion of when a theory adds expressivity to its subtheory is developed, however it can easily be seen that any model-theoretically conservative and finitely axiomatizable theory of truth over PA will witness this criterion. Let us summarize it in here. Authors treat expressiveness as being a property of a language rather than a theory. However, they notice that it makes little sense to take just a set of formulae as formally representing a concept. This leads to a relativized notion of expressiveness: Fischer and Horsten talk about the language L₁ being more expressive with respect to a theory T₁ than a language L₂ with respect to a theory T₂ (with L₂ ⊆ L₁, T₂ ⊆ T₁), which, by definition, means that there exists a class of models of T₁ which is L₁-elementary (i.e., there is an L₁ sentence φ such that the models in this class are precisely the T₁ models satisfying φ) but the class consisting of reducts of these models is not L₂ elementary. Philosophical motivation behind this definition is quite clear: if L₁, T₁, L₂, T₂ satisfy the defining condition then L₁ has the resources to single out a class of models that cannot be isolated by means provided by L₁. Then in order to witness the expressive function of truth authors take as L₂ the language of arithmetic, as T₂ – Robinson’s arithmetic Q, as L₂ the language L₁ augmented with a fresh unary predicate T, and finally as T₁ – Q extended with all Tarski biconditionals for sentences of L₁. Let us sketch the argument: in [6] it was argued that the theory PT⁻⁺ INT↾tot (PT⁻ is authors’ notation) axiomatizes the class of models whose L₁ reducts are precisely all models of PA (i.e., it is semantically conservative over PA) and is finitely axiomatizable. The class of all models of Q which are models of PA is not L₁-elementary, since PA is not finitely axiomatizable. Since, as shown in [4] PT⁻⁺ INT↾tot is not semantically conservative this argument breaks down. However, it can be easily fixed (even with the use of the same theory of truth), since no extension of PA in the arithmetical language is finitely axiomatizable (PA is essentially reflexive).
2. Form$_{L_{PA}}(x)$, Form$_{L_{PA}}^{<1}(x)$, Form$_{L_{PA}}^{v}(x)$, and Sent$_{L_{PA}}(x)$ are natural arithmetical formulae strongly representing in PA the sets of (Gödel codes of) formulae of $L_{PA}$, formulae of $L_{PA}$ with at most one free variable, formulae of $L_{PA}$ in which $v$ is the unique free variable, and sentences of $L_{PA}$, respectively. Likewise, Term$_{L_{PA}}$ and ClTerm$_{L_{PA}}$ strongly represent the set of arithmetical terms and closed terms, respectively.

3. Subst$(\phi, t, \psi)$ is an arithmetical formula strongly representing in PA the relation “$\psi$ is the effect of substituting the term $t$ for every variable in the formula $\phi$.”

4. If $\tau$ is a $L_{PA}$ formula or a term then $\Gamma \tau$ denotes either its Gödel code or the numeral denoting the Gödel code of $\tau$ (context-dependently). These are the unique ways of using $\Gamma \tau$ in this paper.

5. $x$ denotes the (Gödel code of the) standard numeral for $x$, i.e., $\Gamma S \ldots S(0)$.

6. $y^\circ$ is the standard arithmetically definable function representing the value of term (coded by) $y$. For example, in any model of PA the formula $x = \Gamma (S(0) + S(S(0)) \times S(S(S(0))))^\circ$ is satisfied only by $x$ equal to 7.

7. By using variables $\phi, \psi$ we implicitly restrict quantification to (Gödel codes of) arithmetical sentences. For example, by $\forall \phi \Psi(\phi)$ we mean $\forall x \left( \text{Sent}_{L_{PA}}(x) \rightarrow \Psi(x) \right)$ and by $\exists \phi \Psi(\phi)$ we mean $\exists x \left( \text{Sent}_{L_{PA}}(x) \land \Psi(x) \right)$. For brevity, we will sometimes also use variables $\phi, \psi$ to range over arithmetical formulae, whenever it is clear from the context which one we mean. Similarly,

(a) $\phi(v), \psi(v)$ range over arithmetical formulae with at most one indicated free variable (i.e., $\phi(v)$ is either a formula with $v$ being its unique free variable or $\phi(v)$ has no free variables); $\phi(x), \psi(x), \ldots$ range over arbitrary arithmetical formulae;

(b) $s, t$ range over codes of closed arithmetical terms;

(c) $v, u, v_1, v_2, \ldots, w, w_1, w_2, \ldots$ range over codes of variables.

8. To enhance readability we suppress the formulae representing the syntactic operations. For example, we write $\Phi(\psi \land \eta)$ instead of $\exists x \left( \Phi(x) \land "x is the conjunction of \psi and \eta" \right)$,

similarly, we write $\Phi(\psi(t))$ instead of $\exists x \left( \Phi(x) \land x = \text{Subst}(\psi, t) \right)$.

The main objective of this study is to measure the strength of theories that are compositional, but do not enjoy the global axiom for commutativity with the negation, i.e.,

$$\forall \phi \left( T(\neg \phi) \equiv \neg T(\phi) \right).$$ (NEG)

Let us formulate the theories which will be of the main interest.

**Definition 1.6.** $PT^-$ is the axiomatic truth theory with the following axioms for the truth predicate:
In the arithmetized language we treat $\wedge$ and $\forall$ as symbols defined contextually, i.e., $
abla \wedge \psi = \neg(\neg\phi \lor \neg\psi)$ and $\forall \phi = \neg\exists v\neg \phi$. Then it is easy to check that the following sentences are provable in PT$^-$:

1. $\forall \phi, \psi (T(\phi \land \psi) \equiv (T(\phi) \land T(\psi)))$.
2. $\forall \phi, \psi (T(\neg(\phi \land \psi)) \equiv (T(\neg\phi) \lor T(\neg\psi)))$.
3. $\forall v \forall \phi(v) (T(\forall v \phi) \equiv (\forall x T(\phi(x))))$.
4. $\forall v \forall \phi(v) (T(\neg\forall v \phi) \equiv (\exists x T(\neg\phi(x))))$.

In PT$^-$ the internal logic (i.e., the logic of all true sentences) is modelled after the Strong Kleene Scheme.$^{14}$ Let us observe that axioms of PT$^-$ make it possible to accept a disjunction $\phi \lor \psi$ as true simply on the basis of the truth of one of $\phi$ and $\psi$ and regardless of whether the second one has its truth value determined. The second theory we will study is more cautious in this respect. Let us define

$$\text{tot}(\phi(v)) := \text{Form}^1_{\mathbb{PA}}(\phi(v)) \land \forall x (T(\phi(x)) \lor T(\neg\phi(x))).$$

In particular, if $\psi$ is a sentence, then

$$\text{PAT}^- \vdash \text{tot}(\psi) \equiv (T(\psi) \lor T(\neg\psi)),$$

where PAT$^-$ is the extension of PA in $L_{\mathbb{PA}} \cup \{T\}$, with no nonlogical axioms for $T$.

**Definition 1.7.** WPT$^-$ is the axiomatic truth theory with the following axioms for the truth predicate:

1. (a) $\forall s, t (T(s = t) \equiv (s^\circ = t^\circ))$
   (b) $\forall s, t (T(\neg s = t) \equiv (s^\circ \neq t^\circ))$
2. (a) $\forall \phi, \psi (T(\phi \lor \psi) \equiv (\text{tot}(\phi) \land \text{tot}(\psi) \land (T(\phi) \lor T(\psi))))$
   (b) $\forall \phi, \psi (T(\neg(\phi \lor \psi)) \equiv T(\neg\phi) \land T(\neg\psi))$
3. (a) $\forall v \forall \phi(v) (T(\exists v \phi(v)) \equiv \text{tot}(\phi(v)) \lor \exists x T(\phi(x)))$
   (b) $\forall v \forall \phi(v) (T(\neg\exists v \phi) \equiv \forall x T(\neg\phi(x)))$
4. $\forall \phi (T(\neg\phi) \equiv T(\neg\phi))$
5. $\forall v \forall \phi(v) \forall s, t (s^\circ = t^\circ \rightarrow T(\phi(s)) \equiv T(\phi(t)))$.

Using the abovementioned conventions regarding $\wedge$ and $\forall$, it is an easy exercise to show that the following sentences are provable in WPT$^-:$

---

$^{14}$ A discussion of Strong and Weak Kleene logics can be found for instance in [11]. Moreover, the introduction to [10] provides intuitions concerning the two logics.
MODELS OF POSITIVE TRUTH 151

1. \(\forall \phi, \psi \ (T(\phi \land \psi) \equiv (T(\phi) \land T(\psi))).\)

2. \(\forall \phi, \psi \ (T(\neg (\phi \land \psi)) \equiv (\text{tot}(\phi) \land \text{tot}(\psi) \land (T(\neg \phi) \lor T(\neg \psi))).)\)

3. \(\forall v \forall \phi(v) \ (T(\forall v \phi) \equiv (\forall x \ T(\phi(x))).)\)

4. \(\forall v \forall \phi(v) \ (T(\neg \forall v \phi) \equiv (\text{tot}(\phi(v)) \land \exists x \ T(\neg \phi(x))).)\)

In WPT\(^-\) the internal logic is modelled after the Weak Kleene Scheme. (W)PT\(^-\) can be seen as a natural stratified counterpart of (W)KF\(^-\).\(^{15}\) Since in particular (W)PT\(^-\) is a subtheory of (W)KF\(^-\) and the latter is well known to be model-theoretically conservative over PA (see [1]; we will outline a direct proof of model-theoretical conservativity of PT\(^-\) in \(\S 3\)), we have

**PROPOSITION 1.8.** PT\(^-\) and WPT\(^-\) are model-theoretically conservative.

In particular we see that the axiom (NEG) may contribute to the strength of truth theories: it is easy to see that (W)PT\(^-\) + (NEG) is deductively equivalent to the theory CT\(^-\), hence in particular by the well-known theorem of Lachlan (see [12, 18]) it is not semantically conservative.

For the sake of convenience let us isolate one easily noticeable feature of PT\(^-\) and WPT\(^-\):

**DEFINITION 1.9 (UTB).** Let \(\phi(x_0, \ldots, x_n)\) be any arithmetical formula. UTB\(^-\)(\(\phi\)) is the following \(L_T\) sentence

\[
\forall t_0 \ldots t_n \ (T(\upharpoonright \phi(t_0, \ldots, t_n)) \equiv \phi(t_0^\circ, \ldots, t_n^\circ)).
\]

\((\text{UTB}^- (\phi))\)

Define

\[\text{UTB}^- := {\text{UTB}^-(\phi(x_0, \ldots, x_n)) \mid \phi(x_0, \ldots, x_n) \in \mathcal{L}_{PA}}.\]

Define UTB to be the extension of UTB\(^-\) with all instantiations of induction scheme with formulae containing the truth predicate. Note that when we treat UTB as an actual theory the above scheme is taken over standard \(\phi\) not their Gödel codes in some model of arithmetic.

**FACT 1.10.** Both PT\(^-\) and WPT\(^-\) prove UTB\(^-\).

As we already noted, in [7, 8, 9] (this last philosophical motivation was summarized also in [4]) authors motivated the need for a weak theory of truth which would be able to prove in a single sentence the fact that every arithmetical formula satisfies the induction scheme. Such a fact can be naturally expressed by a sentence

\[
\forall v \forall \phi(v) \left[ (\forall x \ (T(\phi(x)) \rightarrow T(\phi(x+1))) \rightarrow (T(\phi(0)) \rightarrow \forall x \ T(\phi(x)))) \right].
\]

\((\text{INT})\)

Note that according to our Convention 1.5 the variable \(\phi(v)\) ranges over arithmetical formulae. In other words the axiom INT states that all arithmetical formulae satisfy the induction scheme. For further usage let us abbreviate the formula

\[
(\forall x \ (T(\phi(x)) \rightarrow T(\phi(x+1))) \rightarrow (T(\phi(0)) \rightarrow \forall x \ T(\phi(x)))
\]

by INT(\(\phi(v))\). Using Fact 1.10 we see that both PT\(^-\) + INT and WPT\(^-\) + INT can prove any arithmetical instance of the induction schema in a uniform way, for each formula using the same finitely many axioms.\(^{16}\) In particular, it can be finitely axiomatized by taking

\(^{15}\) For the definition of all mentioned theories not defined in this paper consult [12] or [11] (for WKF).

\(^{16}\) The proof is really easy: we fix \(\phi(x)\) (with parameters), prove the instantiation of the UTB\(^-\)(\(\phi\)) scheme for \(\phi\), and substitute \(\phi(x)\) for \(T(\phi(x))\) in INT.
I$_1$ together with axioms for the truth predicate from P$^-$ and (INT) (we will outline the argument in §4, Theorem 4.5). To achieve this goal, however, none of the discussed theories use the full strength of (INT). By UTB$^-$ every standard formula is total, provably in WPT$^-$. Hence, it makes good sense to consider a version of (INT) restricted to total arithmetical formulae, i.e.,

$$\forall \phi \forall \phi(v) \left[ \text{tot}(\phi(v)) \rightarrow \text{INT}(\phi(v)) \right].$$

The theory P$^- + \text{INT} \mid_{\text{tot}}$ was claimed to be model-theoretically conservative in [6] (and then used in [7, 8, 9] as such). However, as shown in [4], the proof of its conservativity contained an essential gap, and no prime model of PA$^+$ admits an expansion to a model of P$^- + \text{INT} \mid_{\text{tot}}$. Moreover, it was shown that every recursively saturated model of PA can be expanded to a model of this theory. In particular, P$^- + \text{INT} \mid_{\text{tot}}$ is model-theoretically stronger than P$^-$ and weaker than UTB and CT$^-$. 

1.5. Summary of the results and the big picture. In the current study we further approximate the class of models expandable to P$^- + \text{INT} \mid_{\text{tot}}$ and compare the strength of UTB with the strength of P$^- + \text{INT}$. Moreover, we show that WPT$^- + \text{INT}$ is model theoretically conservative and meets the requirements posed in [9]. Our results jointly with some known facts from the literature give the following picture of interdependencies between proof-theoretically conservative theories of truth:

where $\rightarrow$ stands for $\leq_M$ and $\Rightarrow$ for $\leq_M$. The inequality UTB $\leq_M$ CT$^-$ was demonstrated in [21], where also the strict inequality TB $\leq_M$ UTB was shown. We sharpen the latter result by showing that P$^- + \text{INT} \mid_{\text{tot}}$ (strictly) separates both theories. The question whether any of $\Rightarrow$ arrows is in fact an $\rightarrow$ arrow is open. Similarly, the relation between classes of models of CT$^-$ and P$^- + \text{INT}$ is unknown.

---

17 Recall that a prime model of PA is a model in which every element is definable without parameters by an arithmetical formula. Every complete extension of PA in the language of arithmetic has exactly one prime model. It has no elementary submodels satisfying PA and thus is minimal. For the definitions and discussion of all notions from the model theory of PA see [18].
§2. Models of $\text{PT}^- + \text{INT} \upharpoonright_{\text{tot}}$. In the paper [4] it has been shown that $\text{PT}^- + \text{INT} \upharpoonright_{\text{tot}}$ is not semantically conservative over PA and, moreover, any (not necessarily countable) recursively saturated model of PA admits an expansion to a model of $\text{PT}^- + \text{INT} \upharpoonright_{\text{tot}}$. The nonconservativity result was obtained by demonstrating that no prime model of PA can be expanded to a model of that theory. Now, we will show a strengthening of that result. Let us first recall one definition.

**Definition 2.1.** Let $M$ be a model of PA. We say that $M$ is recursively saturated if any computable type (with finitely many parameters from $M$) $p(x) = \{\phi_i(x) \mid i \in \mathbb{N}\}$ is realised in $M$. We call it short recursively saturated if the same holds for every computable type $p$ satisfying the additional restriction that it contains a formula $x < a$ for some parameter $a$.

In other words a model is short recursively saturated if it realises all types which are finitely realised below some fixed element. This notion is strictly weaker than full recursive saturation. For example, the standard model $\mathbb{N}$ is short recursively saturated, but not recursively saturated. More generally, a countable model is short recursively saturated if and only if it has a recursively saturated elementary end extension, see [26], Theorem 2.8.

**Theorem 2.2.** Let $M \models \text{PA}$ and suppose that $M$ has an expansion $(M, T) \models \text{PT}^- + \text{INT} \upharpoonright_{\text{tot}}$. Then $M$ is short recursively saturated.

The proof of our theorem will closely parallel the proof of Theorem 4 from [4]. In particular, we will again use a propositional construction invented by Smith.

**Definition 2.3.** Let $(\alpha_i)_{i \leq c}, (\beta_i)_{i \leq c}$ be any sequences of sentences. We define the disjunction with stopping condition $(\alpha_i)$, $c, a \bigvee_{i=i_0}^{c, a} \beta_i,$ by backwards induction on $i_0$ as follows:

1. $\bigvee_{i=i_0}^{c,a} \beta_i = \alpha_c \land \beta_c$.
2. $\bigvee_{i=i_0}^{c,a} \beta_i = \neg (\alpha_k \land \neg \beta_k) \land \left( (\alpha_k \land \beta_k) \lor (\neg \alpha_k \land \bigvee_{i=k+1}^{c,a} \beta_i) \right)$.

We may think that this is a formalization in propositional logic of the following instruction: for $i$ from $i_0$ up to $c$, search for the first number $j$ such that $\alpha_j$ holds and then check whether also $\beta_j$ holds. Then stop your search. The whole formula is true if this $\beta_j$ is true and is false if either $\beta_j$ is false or there is no $j$ such that $\alpha_j$ holds. It turns out that this intuition may be partially recovered in theories of truth, even if one does not assume that the truth predicate satisfies induction axioms.

**Lemma 2.4.** Fix $(M, T) \models \text{PT}^-$. Suppose that $(\alpha_i)_{i \leq c}, (\beta_i)_{i \leq c}$ are sequences of arithmetical sentences coded in $M$. Suppose that the least $j$ such that $T(\alpha_j)$ holds is standard, say $j = j_0$, and that for any $k \leq j_0$ either $T(\beta_k)$ or $T(\neg \beta_k)$ holds. Then

1. $(M, T) \models T \left( \bigvee_{i=0}^{c,a} \beta_i \right) \equiv T (\beta_{j_0}).$
2. $(M, T) \models T \left( \neg \bigvee_{i=0}^{c,a} \beta_i \right) \equiv T (\neg \beta_{j_0}).$

---

18 This simply means that $M \models \text{"a, } \beta \text{ are sequences whose all elements are arithmetical sentences."}
For a proof, see [4], Lemma 2.3. Now we are ready to prove that any model of PA expandable to a model of PT $\models_{\text{tot}}$ is short recursively saturated.

Proof. Fix any computable type $p(x) = (\phi_i(x) \land x < a)_{i \in \omega}$ (with a parameter $a$) and suppose that for any finite set $\phi_0, \ldots, \phi_k$ there is some $b_k < a$ such that $M \models \phi_0(b_k) \land \cdots \land \phi_k(b_k)$. Let
\[
\begin{align*}
a_0(x) &= \neg x < a \lor \neg \phi_0(x) \\
a_{j+1}(x) &= x < a \land \phi_0(x) \land \cdots \land \phi_j(x) \land \neg \phi_{j+1}(x).
\end{align*}
\]
In a sense, formulae $a_j(x)$ measure how much of the type $p$ is realised by $x$. Now, if the type $p$ is omitted in the model $M$, then for any $x$, there exists a standard $j$ such that $(M, T) \models Ta_j(x)$. Let $\beta_j(y)$ be defined as
\[
\beta_j(y) = y < a \land \phi_0(y) \land \cdots \land \phi_{j+1}(y).
\]
Now, fix any nonstandard $c$ and consider the (nonstandard) formula
\[
\phi(x, y) = \bigvee_{i=0}^{c, a(x)} \beta_i(y).
\]
By Lemma 2.4 and our assumption that the type $p$ is omitted in $M$ the sentence $\phi(x, y)$ is either true or false for any fixed $x, y \in M$. But this means that the formula $\phi(x, y)$ is total. One can check that then the formula
\[
\psi(z) = \exists y \forall x < z \phi(x, y)
\]
is also total. Note that this formula intuitively says that there is a $y$ which satisfies more of a type $p$ than any of the elements of $M$ up to $z$. Now we will show that $\psi(z)$ is progressive, i.e.,
\[
(M, T) \models \forall z \left(T\psi(z) \rightarrow T\psi(z+1)\right).
\]
Fix any $z$ and suppose that $T\psi(z)$ holds. Then there exists a $y$ such that $T(\forall x < z \phi(x, y))$. Now let $j$ be the least number such that $Ta_j(z)$. Since $j$ is a standard number and $p$ is a type there exists $y'$ such that $\phi_0(y') \land \cdots \land \phi_{j+1}(y')$ holds in $M$, i.e., $(M, T) \models T\beta_j(y')$. Let $y'' = y$ if also
\[
T\left(\phi_0(y') \land \cdots \land \phi_{j+1}(y')\right)
\]
and $y'' = y'$ otherwise. In other words we fix either $y$ or $y'$, whichever satisfies “more” formulae $\phi_i$. One readily checks that then
\[
(M, T) \models \forall x < z + 1 \phi(x, y'').
\]
We have shown that the formula $\psi(z)$ is total and progressive. By the internal induction for total formulae this means that
\[
(M, T) \models \forall z \ T\psi(z).
\]
In particular, we have $T\psi(a)$, where $a$ is the parameter used as a bound in the type $p$. But then for some $d$, we have
\[
(M, T) \models \forall x < a \ T\phi(x, d).
\]
Now, since $p$ is a type, for an arbitrary $k \in \omega$, there exists some $x < a$ such that $\phi_0(x) \land \cdots \land \phi_k(x)$. Since $(M, T) \models T\phi(x, d)$ it follows that $d < a \land \phi_0(d) \land \cdots \land \phi_{k+1}(d)$. As
we have chosen an arbitrary \( k \) we see that actually \( d \) satisfies the type \( p \). We conclude that \( M \) is short recursively saturated.  

Let us summarize our findings from [4] and this paper:

- Any recursively saturated model of PA (possibly uncountable) admits an expansion to a model of PT\(^-\) + INT \( \restriction \) tot.
- If a model \( M \) expands to a model of PT\(^-\) + INT \( \restriction \) tot then it is short recursively saturated.

Unfortunately, we do not know whether any of the implications reverses.

Cieśliński and Engström have (independently) found the following characterization of the class of models of PA which admit an expansion to a model of TB, i.e., the truth theory axiomatized with the induction scheme for the whole language and the following scheme of Tarski’s biconditionals:

\[ T(⌜\phi⌝) \equiv \phi, \]

where \( \phi \) is an arithmetical sentence.

**Theorem 2.5** (Cieśliński, Engström\(^{19}\)). Let \( M \) be a nonstandard model of PA. Then, the following are equivalent:

1. \( M \) admits an expansion to a model \( (M, T) \models TB \).
2. There exists an element \( c \in M \) such that for all (standard) arithmetical sentences \( \phi, M \models \Gamma \phi \in c \) iff \( M \models \phi \), i.e., \( M \) codes its own theory.

It can be easily shown that every nonstandard short recursively saturated model \( M \models PA \) satisfies the second item of the above characterization. Hence, every short recursively saturated model of PA admits an expansion to a model of TB. Thus, we obtain the following corollary:

**Corollary 2.6.** TB \( \leq_M \) PT\(^-\) + INT \( \restriction \) tot, i.e., every model \( M \models PA \) which admits an expansion to a model of PT\(^-\) + INT \( \restriction \) tot also admits an expansion to a model of TB.

2.1. A non-result. In [4], it was shown that every recursively saturated model of PA can be expanded to a model of PT\(^-\) + INT \( \restriction \) tot. The proof actually showed that if \( M \models PA \) is a recursively saturated model, then a certain positive operator \( \Gamma : M \rightarrow M \) reaches its fixpoint \( T \) exactly in \( \omega \) steps. The operator \( \Gamma \) will be defined below. The reader may think of it as a fine-grained version of the Kripke jump operator in which we accept new sentences as true step by step: one connective, quantifier or one direct application of the truth predicate at a time. From this, we could deduce that such a \( T \) is an interpretation for a truth predicate satisfying PT\(^-\) + INT \( \restriction \) tot.

One could hope to use this method to obtain an analogous result for a bigger class of models of PA. One possible approach would be simply to check whether the operator \( \Gamma \) reaches its fixpoint after \( \omega \) steps under some milder assumptions on \( M \). Below, we show that this is impossible. The discussed operator reaches fixpoints after \( \omega \) steps exactly in recursively saturated models.

**Convention 2.7.** If \( M \models PA \), then Sent\(_{PA}(M)\), Form\(_{\leq 1}^{=1}(M)\), and CTerm\(_{\leq 1}(M)\) denote the set of sentences of \( L_{PA} \), the set of formulae with at most one free variable, and

\(^{19}\) See [3], Theorem 7.
the set of closed terms, respectively, in the sense of \( \mathcal{M} \). In other words, they denote the sets defined by the respective formulae in the model \( \mathcal{M} \).

The definition below is essentially the same as in [12].

**Definition 2.8.** Let \( \mathcal{M} \models \text{PA} \). Fix an arbitrary set \( A \subseteq \mathcal{M} \). We define (externally), a second order formula \( \Theta_{\mathcal{M}} \) with a parameter \( A \):

\[
\Theta_{\mathcal{M}}(\phi, A) := \exists s, t \in \text{CTerm}_{\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (s = t) \land s^2 = t^2\right] \\
\lor \exists s, t \in \text{CTerm}_{\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = \neg(s = t) \land s^2 \neq t^2\right] \\
\lor \exists \psi \in \text{Sent}_{\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (\neg\neg\psi) \land \psi \in A\right] \\
\lor \exists \psi_1, \psi_2 \in \text{Sent}_{\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (\psi_1 \lor \psi_2) \land ((\psi_1 \in A) \lor (\psi_2 \in A))\right] \\
\lor \exists \psi_1, \psi_2 \in \text{Sent}_{\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (\neg(\neg\psi_1 \lor \neg\psi_2) \land (\neg\psi_1 \in A) \lor (\neg\psi_2 \in A))\right] \\
\lor \exists \psi(x) \in \text{Form}^{\leq\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (\exists x \psi) \land \exists x \in \mathcal{M} (\psi(x) \in A)\right] \\
\lor \exists \psi(x) \in \text{Form}^{\leq\mathcal{L}_{\text{pa}}} (\mathcal{M}) \left[M \models \phi = (\neg\exists x \psi) \land \forall x \in \mathcal{M} (\neg\psi(x) \in A)\right].
\]

Let \( \Gamma^{\mathcal{M}} : \mathcal{P}(\mathcal{M}) \rightarrow \mathcal{P}(\mathcal{M}) \) be the function defined as follows:

\[ \Gamma^{\mathcal{M}}(A) = \{ \phi \in \mathcal{M} \mid \Theta_{\mathcal{M}}(\phi, A) \}. \]  

(\( \Gamma \))

Let us now define:

\[
\Gamma^{\mathcal{M}}_0 = \Gamma^{\mathcal{M}}(\emptyset) \\
\Gamma^{\mathcal{M}}_{\alpha+1} = \Gamma^{\mathcal{M}}(\Gamma^{\mathcal{M}}_{\alpha}) \\
\Gamma^{\mathcal{M}}_{\beta} = \bigcup_{\alpha < \beta} \Gamma^{\mathcal{M}}_{\alpha}, \text{ for } \beta \text{ a limit ordinal.}
\]

It can be checked that for some ordinal \( \alpha \) we must get \( \Gamma^{\mathcal{M}}_{\alpha+1} = \Gamma^{\mathcal{M}}_{\alpha} \), i.e., \( \Gamma^{\mathcal{M}}_{\alpha} \) is a fixpoint of \( \Gamma^{\mathcal{M}} \). In general, if \( A \) is any fixpoint of \( \Gamma^{\mathcal{M}} \), then

\[ (\mathcal{M}, A) \models \text{PT}^- \].

Let \( \alpha_{\mathcal{M}} \) denote the least ordinal \( \alpha \) such that \( \Gamma^{\mathcal{M}}_{\alpha} \) is a fixpoint of \( \Gamma^{\mathcal{M}} \).

In [4], the following lemmata were proved:

**Lemma 2.9.** If \( \mathcal{M} \models \text{PA} \) is recursively saturated, then \( \alpha_{\mathcal{M}} = \omega \).

**Lemma 2.10.** If \( \mathcal{M} \models \text{PA} \) and \( \alpha_{\mathcal{M}} = \omega \), then \( (\mathcal{M}, \Gamma^{\mathcal{M}}_{\omega}) \models \text{PT}^- + \text{INT} \upharpoonright_{\text{tot}} \).

Now we shall show that the converse of Lemma 2.9 holds. In particular, our method of finding expansions satisfying \( \text{PT}^- + \text{INT} \upharpoonright_{\text{tot}} \) works only for recursively saturated models.

**Lemma 2.11.** For every nonstandard \( \mathcal{M} \models \text{PA} \), if \( \alpha_{\mathcal{M}} = \omega \), then \( \mathcal{M} \) is recursively saturated.

**Proof.** We prove the contraposition: suppose that a nonstandard model \( \mathcal{M} \) is not recursively saturated.

Let \( p(x) \) be a computable type using parameters from \( \tilde{a} \) which is omitted in \( \mathcal{M} \). Let \( (\phi_i(x, \tilde{y}))_i \) be an arithmetically representable enumeration of formulae in \( p(x) \). Without
loss of generality assume that \( \phi_0(x, \bar{y}) = (x = x) \). Let

\[
\psi_i(x, \bar{y}) = \bigwedge_{j<i} \phi_j(x, \bar{y}) \land \neg \phi_i(x, \bar{y}).
\]

Then, every \( b \in M \) satisfies exactly one of \( \psi_i(x, \bar{a}) \) (since \( p(x) \) is omitted).

Now, for every \( n \in \omega \) we shall define formulae \( \theta_n(x) \) as follows:

\[
\theta_n^0(x) = (x \neq x)
\]
\[
\theta_n^{k+1}(x, \bar{y}) = \psi_{n-(k+1)}(x, \bar{y}) \lor \theta_n^k(x, \bar{y})
\]
\[
\theta_n(x, \bar{y}) = \theta_n^0(x, \bar{y}).
\]

Let us observe that the above construction can be arithmetized and therefore for some \( b \in M \setminus \mathbb{N} \) there exists a (code of a) formula \( \theta_b(x, \bar{y}) \), which is of the following form:

\[
(\psi_0(x, \bar{y}) \lor (\psi_1(x, \bar{y}) \lor (\psi_2(x, \bar{y}) \lor \ldots (\psi_{b-1}(x, \bar{y}) \lor x \neq x) \ldots)).
\]

Then, for each \( c \in M \) there exists \( n \in \omega \) such that \( \theta_b(c, \bar{a}) \in \Gamma^n_\infty \), since each \( c \) satisfies some \( \psi_i(x, \bar{a}) \). But also for every \( i \in \omega \) there exists \( c \in M \) such that the unique \( n \) for which \( \psi_n(c, \bar{a}) \) holds is greater than \( i \). Consequently, there is no \( k \in \omega \) for which

\[
\theta_b(c, \bar{a}) \in \Gamma^M_k
\]

for every \( c \in M \). Indeed, fix \( k \) and pick \( c, n \) such that \( \psi_n(c, \bar{a}) \) holds and \( n > k \). Then \( \theta_b(c, \bar{a}) \notin \Gamma^M_k \), for otherwise, using the fact that \( \psi_i(c, \bar{a}) \) is false for \( i < n \), by induction on \( i \) we would show that

\[
\theta_{b-i}(c, \bar{a}) \in \Gamma^M_{k-i},
\]

for every \( i \leq k \). The above condition is clearly false for \( i = k \).

It follows that \( \forall v \exists \theta(v, \bar{a}) \notin \Gamma^M_\omega \) and consequently the PT\(^-\) axiom

\[
\forall v \forall \phi(v) \left( T(\forall v \phi) \equiv \forall x \ T(\phi(x)) \right)
\]

is not satisfied in \( (M, \Gamma^M_\omega) \). Hence \( \alpha_M \neq \omega \). \( \square \)

§3. Models of PT\(^-\) + INT. Since we have shown that PT\(^-\) + INT \( |_{\text{tot}} \) is not a model-theoretically weak theory, as was originally hoped, one could start wondering whether it differs in some significant respect from PT\(^-\) + INT. In this section we will show that actually this is the case. Namely, it turns out that PT\(^-\) + INT is still model-theoretically stronger than PT\(^-\) + INT \( |_{\text{tot}} \). As we shall see any model of PA expandable to a model of PT\(^-\) + INT is also expandable to a model of UTB. We know that any model of PA expandable to a model of UTB is recursively saturated and that this containment is strict, i.e., not every recursively saturated model of PA admits an expansion to a model of UTB.\(^{20}\)

\(^{20}\) We know that there exist rather classless recursively saturated models of PA, i.e., recursively saturated models \( M \models PA \) with the following property: for every \( X \subseteq M \) such that every initial segment of \( X \) is coded in \( M \), the set \( X \) is definable in \( M \) with an arithmetical formula (with parameters). Since no subset of \( M \) definable with an arithmetical formula can satisfy UTB\(^-\) we see that no such model \( M \) can admit an expansion to a model of UTB. The existence of recursively saturated, rather classless models has been shown by Kaufmann in [17] under an additional set-theoretic assumption \( \diamond \). The assumption has been dropped by Shelah, [25], Application C, p. 74.
On the other hand, it has been shown in [4], Theorem 3.3 that any recursively saturated model of PA admits an expansion to a model of $\text{PT}^{-} + \text{INT} \upharpoonright \text{tot}$.

**Theorem 3.1.** Suppose that $(M, T)$ is a model of $\text{PT}^{-} + \text{INT}$. Then there exists a $T'$ such that $(M, T') \models \text{UTB}$.

*Proof.* Let $(M, T) \models \text{PT}^{-} + \text{INT}$. We will find $T'$ such that $(M, T') \models \text{UTB}$. Without loss of generality we may assume that $M$ is nonstandard. As in the previous section we will use Lemma 2.4. Let us fix any primitive recursive enumeration $(\phi_i)_{i=0}^{\infty}$ of arithmetical formulae. Then, let

$$\alpha'_i(\phi, t)$$

be defined as the (formalized version of the) formula “$t$ is a (finite) sequence of terms $(t_1, \ldots, t_n)$ and $\phi = \phi_i(t_1, \ldots, t_n)$” and let

$$\alpha_i(\phi, t, b) = \alpha'_i(\phi, t) \lor \bar{i} > b.$$ 

Let

$$\beta'_i(t)$$

be defined as “$t$ is a (finite) sequence of terms $t_1, \ldots, t_n$ and $\phi_i(t_1, \ldots, t_n)$.” Let

$$\beta_i(t, b)$$

be $\beta'_i(t) \land \bar{i} \leq b$. Note that $\phi$ is not a free variable of the formula $\beta_i$. Let us fix any nonstandard $c \in M$ and let

$$\tau(\phi, t, b) = \bigvee_{i=0}^{c,\alpha(\phi,t,b)} \beta_i(t).$$

Note that for any standard $c$ the predicate $\tau$ is equivalent to the very simple arithmetical truth predicate:

$$\tau_n(\phi, t, b) = \bigvee_{i=0}^{n} \phi = \phi_i(t) \land \phi_i(t) \land \bar{i} < b.$$ 

At this point one may wonder what is the role of the variable $b$. It is indeed technical. We artificially truncate our truth predicates so that they work only for the first $b$ formulae. This is to some extent controlled by the parameter $c$ in the definition of $\tau$, since whenever $c$ is standard the formula $\tau$ works like a truth predicate only for the first $c$ sentences. However, $c$ is not a variable in the formula $\tau$, but rather a parameter describing the syntactic shape of $\tau$, whereas we need this truncation to be expressed with a variable for reasons which will shortly become clear.

It turns out that for some parameter $b$ the formula given by

$$T'(\phi) = \exists t \ T(\tau(\phi, t, b))$$

satisfies UTB. We will prove this claim in a series of lemmata. This will obviously conclude our proof. □

**Lemma 3.2.** Let $\tau'(\phi, t) = \tau(\phi, t, b)$ for some fixed nonstandard $b$. Then, for an arbitrary standard arithmetical formula $\phi(v_1, \ldots, v_n)$ and an arbitrary sequence of terms $t = (t_1, \ldots, t_n)$, possibly nonstandard (the length of the sequence is assumed to be standard)

$$(M, T) \models T \tau'(\phi(t_1, \ldots, t_n), t) \equiv \phi(t_1, \ldots, t_n).$$
Proof. If φ is standard, then φ = φi for some standard i. So by Lemma 2.4

\[(M, T) \models T' (\phi(t_1, \ldots, t_n), t) \equiv \beta_i(t) \equiv \phi_i(t_1, \ldots, t_n),\]

which is exactly the claim of the lemma. \[\square\]

Note that the above lemma is true in pure PT−. We have used no induction at all. Now, we only need to check that for some parameter b the predicate T′(ϕ, t) defined as Tτ(ϕ, t, b) is fully inductive.

Lemma 3.3. Let T′ be defined as in the above proof. Then, for some b the formula T′(ϕ, t) = τ(ϕ, t, b) is total and consistent i.e., for all ϕ and t, exactly one of Tτ(ϕ, t), T¬τ(ϕ, t) holds.

Proof. Note first that for any standard b the formula τ(ϕ, t, b) is total and consistent. Namely, since a_i(ϕ, t, n) is true for any i > n, we see that for any ϕ, t the least i such that a_i(ϕ, t, n) holds is standard (it is at most n + 1) and then the assumptions of Lemma 2.4 are satisfied. This implies that for any fixed ξ the formula Tτ(ξ, t, n) is equivalent to some φ_i(t') ∧ i ≤ n, which is a standard formula. This implies that for any t, exactly one of Tτ(ξ, t, n), T¬τ(ξ, t, n) holds.

Now, consider the formula

\[\psi(b) = \forall \phi, t \left( \tau(\phi, t, b) \lor \neg\tau(\phi, t, b) \right).\]

We have just shown that for an arbitrary standard n we have T∀b < n ψ(b). So by internal induction we have for some nonstandard d_1

\[T \left( \forall b \leq d_1 \forall \phi, t \left( \tau(\phi, t, b) \lor \neg\tau(\phi, t, b) \right) \right),\]

which gives

\[\forall b \leq d_1 \forall \phi, t \left( T\tau(\phi, t, b) \lor T\neg\tau(\phi, t, b) \right).\]

Similarly, let

\[\xi(b) = \exists \phi, t \left( \tau(\phi, t, b) \land \neg\tau(\phi, t, b) \right).\]

Suppose that T(∃d < b ξ(d)) holds for any nonstandard b < d_1. Then by underspill we would have Tξ(n) for some n ∈ ω. But we have just shown that this is impossible. So there exists some nonstandard b < d_1 such that for any d ≤ b and any φ, t at most one of Tτ(ϕ, t, d), T¬τ(ϕ, t, d) holds. At the same time we know that at least one of these formulae hold. So T′(ϕ, t) = τ(ϕ, t, b) is total and consistent. \[\square\]

We are very close to showing that we have defined a predicate satisfying full induction. Before we proceed, we have to introduce some new notation. Let η be any formula containing a unary predicate P not in the language of PT− and let ξ(ν) be an arbitrary formula with one free variable. Then by η[ξ/P] (or simply η[ξ]) we mean a formula resulting from substituting ξ(t) for any instance of P(t) in η, where t is an arbitrary term, not necessarily closed. We assume that all the variables in η have been renamed so as to avoid clashes.

Let us give an example. Let η(x, y) = P(x + y) ∧ ∃z (z = y ∧ P(z)). Let ξ(ν) = (ν > 0). Then,

\[\eta[ξ] = x + y > 0 \land \exists z (z = y \land z > 0).\]

Now, basically, we would like to finish the proof in the following way. Let T′ be a total formula defined as in the above lemmata and let η be an arbitrary standard formula...
from the arithmetical language enlarged with a fresh unary predicate $P$. Then, applying compositional axioms a couple of times we see that

$$T(\eta[\tau]) \equiv \eta[T \tau].$$

Let us call this principle the generalized commutativity. If this were true, then we could conclude our proof. Namely, by the internal induction principle we know that

$$(\forall x \ (T(\eta[\tau](x)) \to T(\eta[\tau](x + 1))) \to (T(\eta[\tau](0)) \to \forall x \ T(\eta[\tau](x)))$$

which, by generalized commutativity, would allow us to conclude that

$$(\forall x \ (\eta[T \tau](x) \to (\eta[T \tau](x + 1))) \to (\eta[T \tau](0) \to \forall x \ \eta[T \tau](x)).$$

Since the choice of $\eta$ was arbitrary, this precisely means that $\tau$ satisfies the full induction scheme.

Unfortunately, the generalized commutativity principle in the form stated above does not even quite make sense, since we would have to apply the truth predicate to a formula containing free variables. Therefore, we have to restate it more carefully.

**Definition 3.4.** Fix a unary predicate $P$. Let $\eta$ be an arbitrary formula from the language containing that predicate. We say that $\eta$ is in semirelational form if the predicate $P$ is applied only to variables rather than to arbitrary terms.

We may always assume that formulae we use are semirelational, since we may eliminate any occurrence of $P(t)$ for complex terms $t$, by replacing it with $\exists x \ (x = t \land P(x))$. This is expressed in the following lemma:

**Lemma 3.5.** Any formula is equivalent in first-order-logic to a formula in semirelational form.

Now, we are ready to state generalized commutativity lemma in a proper manner.

**Lemma 3.6.** Let $(M, T) \models PT^-$. Let $T * \xi(x) := T(\xi(x))$ for every $x$. Suppose that $\xi$ is total and consistent. Let $\eta$ be an arbitrary standard formula from the arithmetical language extended with a fresh unary predicate $P$ and assume that $\eta$ is in semirelational form. Then the formula $\eta[\xi]$ is total and consistent, and

$$(M, T) \models \forall x_1, \ldots, x_n \ (T(\eta[\xi](x_1, \ldots, x_n)) \equiv \eta[T * \xi](x_1, \ldots, x_n)).$$

The lemma generalizes to the case, where the predicate $P$ is not unary (i.e., $\xi$ may have more than one variable) as well as to the case with finitely many (total and consistent) formulae $\xi_1, \ldots, \xi_n$. The proof may be easily adapted to cover these cases. We will actually use the lemma for the case with $P$ binary.

**Proof.** We prove both claims simultaneously by induction on the complexity of $\eta$. Suppose that $\eta$ is an atomic formula. Then, it is either of the shape $s = t$ for some standard arithmetical terms $s, t$, or of the form $P(x)$.

In the first case $\eta[\xi] = \eta$ and the following equivalences hold:

$$T(s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n)) \equiv s(x_1, \ldots, x_n)^\circ = t(x_1, \ldots, x_n)^\circ$$

$$= s(x_1, \ldots, x_n) = t(x_1, \ldots, x_n)$$

$$= \eta[T * \xi](x_1, \ldots, x_n).$$

If $\eta = P(x)$, then $\eta[\xi] = \xi$ and

$$T(\eta[\xi](x)) = T\xi(x) = \eta[T * \xi](x).$$
So let us prove the induction step. If $\eta$ is a conjunction or disjunction, then the proof is straightforward (the fact that a conjunction or disjunction of sentences which are either true or false is itself either true or false is an easy application of the compositional axioms of \(PT^-\)). If $\eta = \neg \rho$, then we know by induction hypothesis that $\rho[\eta]$ is total and consistent. Then, by the compositional axiom for double negation for the truth predicate the formula $\neg \rho[\eta]$ is also total and consistent and the following equivalences hold:

\[
T(\neg \rho[\eta](x_1, \ldots, x_n)) \equiv \neg T(\rho[\eta](x_1, \ldots, x_n))
\equiv \neg \rho[T * \eta](x_1, \ldots, x_n).
\]

The induction step for quantifier axioms is also simple. Let us prove it for the existential quantifier. Suppose that $\eta = \exists x \ \rho(x, x_1, \ldots, x_n)$. Then,

\[
T(\exists x \ \rho[\xi](x, x_1, \ldots, x_n)) \equiv \exists x \ T(\rho[\xi](x, x_1, \ldots, x_n))
\equiv \exists x \ \rho[T * \xi](x, x_1, \ldots, x_n)
\equiv \eta[T * \xi](x_1, \ldots, x_n).
\]

The second equivalence follows by the induction hypothesis and the last equality by definition. So let us check that $\eta[\xi]$ is total and consistent. Suppose that $T(\exists x \ \rho[\xi](x, x_1, \ldots, x_n))$ does not hold. Then, by compositional axioms for the truth predicate there is no $x$ such that

\[
T(\rho[\xi](x, x_1, \ldots, x_n)).
\]

By induction hypothesis $\rho$ is total and consistent, so for all $x$ we must have

\[
T(\neg \rho[\xi](x, x_1, \ldots, x_n)).
\]

This entails, again by compositional clauses

\[
T(\neg \exists x \ \rho[\xi](x, x_1, \ldots, x_n)). \quad \square
\]

Now, we are ready to conclude the proof of our theorem.

**Lemma 3.7.** Let $(M, T)$ be any nonstandard model of \(PT^- + \text{INT}\). Suppose that $\tau'(\phi, t)$ satisfies the claim of Lemma 3.3. Then, the predicate $T'(\phi, t)$ defined as $T * \tau'(\phi, t)$ satisfies the full induction scheme.

**Proof.** By internal induction principle the following holds for an arbitrary standard $\eta$ from the arithmetical language extended with one fresh unary predicate $P(v)$:

\[
(\forall x \ (T(\eta[\tau](x))) \rightarrow T(\eta[\tau](x + 1))) \rightarrow \left( T(\eta[\tau](0)) \rightarrow \forall x \ T(\eta[\tau](x)) \right).
\]

Since $\tau'$ is total, if we additionally assume that $\eta$ is semirelational, we can reach the following conclusion by Lemma 3.6:

\[
(\forall x \ (\eta[T * \tau](x) \rightarrow (\eta[T * \tau](x + 1)))) \rightarrow \left( \eta[T * \tau](0) \rightarrow \forall x \ \eta[T * \tau](x) \right).
\]

Since $\eta$ was an arbitrary semirelational formula and any formula is equivalent to a semirelational one, this shows that $T'$ satisfies the full induction scheme. \quad \square

The conclusion of the proof of Theorem 3.1. We have defined a formula $T * \tau'(\phi, t)$ which satisfies full induction scheme and such that for an arbitrary standard $\phi(v_1, \ldots, v_n)$ and an arbitrary sequence of terms $(t_1, \ldots, t_n)$ the following holds:

\[
(M, T) \models T * \tau'(\phi(t_1, \ldots, t_n)) \equiv \phi(t_1^\circ, \ldots, t_n^\circ).
\]
Then, the formula $T'(\phi)$ defined as

$$T \ast \tau'(\phi)$$

satisfies the uniform disquotation axioms of UTB$^-$ as well as the full induction scheme. So, it defines a predicate satisfying UTB in $(M, T)$. □

This model-theoretic result allows us to make some conclusions concerning relative definability of the introduced theories.

**Corollary 3.8.** The theory $\text{PT}^- + \text{INT} \mid_{\text{tot}}$ does not relatively truth define $\text{PT}^- + \text{INT}$.

**Proof.** We have just checked that every model $(M, T) \models \text{PT}^- + \text{INT}$ may be expanded to a model of UTB. As we have already noticed at the very beginning of this section, there exist recursively saturated rather classless models which cannot be expanded to any model of UTB. On the other hand in [4], Theorem 3.3, it has been shown that any recursively saturated model can be expanded to a model of $\text{PT}^- + \text{INT} \mid_{\text{tot}}$. Thus, there exist models of $\text{PT}^- + \text{INT} \mid_{\text{tot}}$ which cannot be expanded to a model of $\text{PT}^- + \text{INT}$. This contradicts relative definability. □

§4. Weak and expressive theories of truth. In this section we shall prove that the theory $\text{WPT}^- + \text{INT}$ meets Fischer-Horsten desiderata. We start by showing that it satisfies the requirements 1–4 (from §1.3).

**4.1. Finite axiomatizability and speed-up.** In the context of theories of truth, which are finite extensions of PA (i.e., which can be axiomatized by axioms of PA and a single sentence) one usually accounts for speed-up and finite axiomatizability by proving that the theory is partially classically compositional and proves a sentence which implies that all standard instances of induction scheme are true. The construction below formalizes this technique. We start with a useful definition:

**Definition 4.1.** Let $\text{CC}(x)$ denote the disjunction of the following formulae

- $\exists s, t \left( x = (s = t) \land (T(x) \equiv s^o = t^o) \right)$
- $\exists \phi, \psi \left( x = (\phi \lor \psi) \land (T(x) \equiv T(\phi) \lor T(\psi)) \right)$
- $\exists \phi \left( x = \neg \phi \land (T(x) \equiv \neg T(\phi)) \right)$
- $\exists \phi \exists v \left( x = \exists v \phi \land (T(x) \equiv \exists y T(\phi(y))) \right)$
- $\text{Form}^{\leq 1}(x) \land \forall s, t \left( s^o = t^o \rightarrow (T(x(s)) \equiv T(x(t))) \right)$

Informally, $\text{CC}(x)$ says that $x$ is a formula on which $T$ behaves compositionally in the sense of classical first-order logic. More specifically, the falsity of $x$ is equivalent to the truth of its negation.

**Definition 4.2.** A truth theory $\text{Th}$ is partially classically compositional if there exists a formula $D(y)$ such that $\text{Th}$ proves the following sentences:

1. $\forall y \left( D(y) \rightarrow \forall x \leq y \ D(x) \right)$;
2. $D(0) \land \forall y \left( D(y) \rightarrow D(y + 1) \right)$;
3. $\forall y \left( D(y) \rightarrow \forall \phi \ (\text{dp}(\phi) \leq y \rightarrow \text{CC}(\phi)) \right)$;

where $\text{dp}(\phi) \leq x$ denotes an arithmetical formula representing the (primitive recursive) relation “the depth of the syntactic tree of $\phi$ is at most $x$.”
If a formula $D$ satisfies the first requirement, we say that it is \textit{downward closed}. If a formula satisfies the second one, we say that it is \textit{progressive}. If a formula $D(y)$ is both downward closed and progressive, we will say that it \textit{defines a cut}. In fact, being \textit{downward closed} is not a very restrictive condition: if $D(y)$ is progressive, then the formula

$$D'(x) := \forall y \leq x \ D(y)$$

defines a downward closed and progressive subset of the universe. The third condition says that if $\phi$ is not too complicated (i.e., its complexity belongs to the cut defined by $D$), then the truth of $\phi$ depends on the truth of its immediate subformulae, according to the rules of Tarski’s semantic for first-order logic. Crucially for our considerations, $\phi$ contradicts neither the law of excluded middle, nor the law of non-contradiction (in the sense of $T$).

**Definition 4.3.** Let $\text{ind}(\phi(x))$ denote the instantiation of the induction scheme with $\phi(x)$, i.e., the universal closure of the following formula:

$$\forall x \ (\phi(x) \rightarrow \phi(x + 1)) \rightarrow (\phi(0) \rightarrow \forall x \ \phi(x)).$$

Following our conventions, we will use $\text{ind}(\cdot)$ to denote an arithmetical formula representing the function which, given a Gödel code of a formula with at most one free variable, returns the Gödel code of the corresponding induction axiom.

**Definition 4.4.** A truth theory proves the truth of induction if there exists a formula $D(y)$ such that $\text{Th}$ proves that $D(y)$ defines a cut and

$$\forall \phi(v) \left( D(\phi(v)) \rightarrow T(\text{ind}(\phi(v))) \right).$$

(T(IND))

We shall say that an $L_T$ theory $\text{Th}$ is finitely axiomatizable modulo $\text{PA}$ if there is a sentence $\phi$ such that the logical consequences of $\text{Th}$ are precisely the logical consequences of $\text{PA} \cup \{\phi\}$. For example, $\text{CT}^-$, $\text{PT}^-$, and $\text{WPT}^-$ are finitely axiomatizable modulo $\text{PA}$ and $\text{TB}$ and $\text{UTB}$ are not.

Now, we have the following theorem whose unique novelty rests on isolating the features that are usually implicitly used to prove finite axiomatizability and speed-up for concrete theories of truth (see e.g., [8]).

**Theorem 4.5.** Assume that

1. $\text{Th}$ is partially classically compositional and proves the truth of induction and
2. $\text{Th}$ is finitely axiomatizable modulo $\text{PA},$

then $\text{Th}$ is finitely axiomatizable and it has super-exponential speed-up over $\text{PA}$.

**Sketch of the proof.** Let $D_1(y)$ define a cut on which $T$ is classically compositional and $D_2(y)$ define a cut on which $\text{Th}$ proves the truth of induction. Then, $D(y) := D_1(y) \land D_2(y)$ defines a cut on which $T$ is classically compositional and proves the truth of induction. Let $\phi$ be a finite-modal-PA axiomatization of $\text{Th}$. Observe that there exists a fixed $k$ such that for every $n$

$$\Sigma_k + \phi \vdash D(n).$$

Indeed, this follows from the fact that

$"D$ is a cut"
is a single $L_T$ sentence provable in Th. Moreover, applying the same reasoning to the sentence:

"$D$ is a cut on which $T$ is classically compositional"

without loss of generality we may assume that $k$ is such that

$$\forall x_0 \ldots \forall x_n \ T(\psi(x_0, \ldots, x_n)) \equiv \psi(x_0, \ldots, x_n)$$

for every $\psi(x_0, \ldots, x_n) \in L_{PA}$. To end our preparations let us assume that $k$ is also large enough to give

$$\Sigma_k + \phi \vdash T(\text{IND}).$$

Now, for every standard formula $\theta(x_0, \ldots, x_n)$ we can prove $\text{ind}(\theta(\bar{x}))$ in $\Sigma_k + \phi$ in the following way:

1. prove that $D$ defines an initial segment on which $T$ is classically compositional;
2. prove $D(\bar{\text{ind}}(\theta(\bar{x})))$ and conclude $D(\bar{\theta}(\bar{x}))$;
3. prove $T(\text{IND})$;
4. using 2. and 3. conclude $T(\bar{\text{ind}}(\theta(\bar{x})))$;
5. prove $\text{UTB}^- (\text{ind}(\theta(\bar{x})))$;
6. conclude $\text{ind}(\theta(\bar{x}))$.

It follows that for some $k$, for every standard $\theta(x)$ there exists a proof of $\text{ind}(\theta(\bar{x}))$ using only the axioms of $\Sigma_k + \phi$, hence the theory

$$\Sigma_k \cup \{\phi\}$$

is a finite axiomatization of Th. To prove that Th has super-exponential speed-up over PA we show that there is a formula $D'(y)$ which provably in Th defines a cut and that

$$\text{Th} \vdash \forall y \ (D'(y) \rightarrow \text{Con}_{PA}(y))$$

where $\text{Con}_{PA}(y)$ is a finitary statement of consistency of PA saying that there is no PA proof of $0 = 1$ which can be coded using less than $y$ bits. For the details, see [7], Theorem 9.

Using the above result we shall show that $\text{WPT}^- + \text{INT}$ is sufficient for satisfying properties 1–3.

PROPOSITION 4.6. $\text{WPT}^-$ is partially classically compositional and $\text{WPT}^- + \text{INT}$ proves the truth of induction.

Proof. Let us define

$$D'(y) := \forall x \leq y \forall \phi \left( \text{dp}(\phi) \leq x \rightarrow (T(\neg \phi) \equiv \neg T(\phi)) \right).$$

Then, it can be easily shown that $D'(y)$ provably in $\text{WPT}^-$ defines a cut on which $T$ is classically compositional (that $D'(y)$ is progressive is assured by the compositional axioms of $\text{WPT}^-$). For convenience, let us define:

$$\text{GC}(x) := \text{Form}(x) \land (T(\text{ucl}(x)) \equiv \forall \sigma \ (\text{Asn}(x, \sigma) \rightarrow T(x[\sigma]))),$$

21 GC stands for “global commutativity.” GC$(x)$ expresses that the truth predicate commutes with the whole block of universal quantifiers in the universal closure of $x$. 

where

1. \( \text{ucl}(\phi(\bar{x})) \) denotes the universal closure of \( \phi(\bar{x}) \);
2. \( \text{Asn}(\phi, \sigma) \) represents the relation “\( \sigma \) is an assignment for \( \phi \)” , i.e., \( \sigma \) is a function defined exactly on the free variables of \( \phi \);
3. \( x[\sigma] \) denotes the result of simultaneous substitution of numerals naming numbers assigned by \( \sigma \) to the free variables of \( x \).

Further define

\[
D(y) := D'(y) \land \forall \phi(\bar{x}) \left( |\text{FV}(\phi(\bar{x}))| \leq y \rightarrow \text{GC}(\phi(\bar{x})) \right),
\]

where \( |\text{FV}(\phi(\bar{x}))| \leq y \) represents the relation “\( \phi(\bar{x}) \) contains at most \( y \) free variables”. For the sake of definiteness we assume that in the quantifier prefix of \( \text{ucl}(\phi) \) the variables are listed in the order of growing indices. It can be easily seen that \( D(y) \) is downward closed. Let us now show that \( D(y) \) is progressive. We work in \( \text{WPT}^- \). Fix an arbitrary \( a \) and suppose that \( D(a) \). Then \( D'(a) \) and, as \( D'(y) \) is progressive, we have also \( D'(a+1) \).

Let us fix an arbitrary formula \( \phi \) with less than \( a + 1 \) free variables and let \( v \) be its free variable with the least index. Then, the following are equivalent

1. \( \text{T(ucl}(\phi)) \)
2. \( \forall x \; \text{T(ucl}(\phi(x/v))) \)
3. \( \forall x \forall \sigma \; (\text{Asn}(\phi(x/v), \sigma) \rightarrow \text{T}(\phi(x/v)[\sigma])) \)
4. \( \forall \sigma \; (\text{Asn}(\phi, \sigma) \rightarrow \text{T}(\phi[\sigma])) \).

The equivalence between 1. and 2. is by the axiom for universal quantifier in \( \text{WPT}^- \). The equivalence between 2. and 3. holds because \( \phi(x/v) \) has \( \leq a \) free variables. The last equivalence holds because each assignment for \( \phi \) consists of an assignment to \( v \) and an assignment to the free variables of \( \phi(x/v) \).

We show that \( \text{WPT}^- + \text{INT} \) proves the truth of induction on \( D(y) \). We work in \( \text{WPT}^- + \text{INT} \). Let us observe that for each formula \( \phi \) we have \( \text{dp}(\phi) \leq \phi \) and \( |\text{FV}(\phi)| \leq \phi. \)

Hence, if \( D(\phi) \), then \( D'(\text{dp}(\phi)) \) and \( D(|\text{FV}(\phi)|) \). In particular, if \( D(\phi) \) then \( T \) is classically compositional on subformulae of \( \phi \) and satisfies generalised commutativity. Let us fix an arbitrary formula \( \phi(v, \bar{w}) \) such that \( D(\phi(v, \bar{w})) \). We have to show \( T(\text{ind}(\phi(v, \bar{w}))) \), i.e.,

\[
T \left( \text{ucl}(\forall v \; (\phi(v) \rightarrow \phi(v+1)) \rightarrow (\phi(0) \rightarrow \forall v \; \phi(v))) \right)
\]

(we skip the reference to \( \bar{w} \) and assume that they are bounded by the universal quantifiers from ucl). Since the formula

\[
(\forall v \; (\phi(v) \rightarrow \phi(v+1)) \rightarrow (\phi(0) \rightarrow \forall v \; \phi(v))
\]

contains less free variables than \( \phi(v) \), we know that (1) is equivalent to

\[
\forall \sigma \; T \left( \forall v \; (\phi(v)[\sigma] \rightarrow \phi(v+1)[\sigma]) \rightarrow (\phi(0)[\sigma] \rightarrow \forall v \; \phi(v)[\sigma]) \right).
\]

(2)

Let us fix an arbitrary \( \sigma \). Then, \( \phi(v)[\sigma] \) is a formula with at most one free variable \( v \). Let us abbreviate it with \( \psi(v) \). Hence, it is enough to show:

\[
T \left( \forall v \; (\psi(v) \rightarrow \psi(v+1)) \rightarrow (\psi(0) \rightarrow \forall v \; \psi(v)) \right).
\]

(3)

\footnote{Being precise, this is a property of our coding. But most natural codings surely have it.}
Since \( \text{dp}(\psi(v)) = \text{dp}(\phi(v)) \) and the depth of (3) is equal to \( \text{dp}(\psi(v)) + 3 \), then \( T \) is classically compositional on (3). Hence (3) is equivalent to
\[
\forall x \ (T(\psi(x)) \rightarrow T(\psi(x + 1))) \rightarrow (T(\psi(0)) \rightarrow \forall x \ T(\psi(x)))
\]
which follows by INT. Hence \( \text{WPT}^- + \text{INT} \) proves the truth of induction on \( D \). \[\square\]

4.2. Formalizing semantical metatheorems: Interpretability of ACA\(_0\). In order to demonstrate that \( \text{PT}^- + \text{INT} \mid_{\text{tot}} \) is capable of formalizing important semantical metatheorems, Fischer and Horsten show that it \( \omega \)-interprets theory ACA\(_0\). We shall show that the Weak Kleene variant of their theory is sufficient for accomplishing this.

PROPOSITION 4.7. There exists an \( \omega \)-interpretation of ACA\(_0\) in \( \text{WPT}^- + \text{INT} \mid_{\text{tot}} \).

In fact, we will do slightly better: we will show that ACA\(_0\) is mutually \( \omega \)-interpretable with \( \text{WPT}^- + \text{INT} \mid_{\text{tot}} \) and with a theory of a progressive truth predicate, which we now introduce:

DEFINITION 4.8. Let CC\((x)\) be the formula introduced in Definition 4.1. Define:
\[
\text{CC}(\langle x \rangle) := \forall \phi \ (\text{dp}(\phi) < x \rightarrow \text{CC}(\phi)).
\]

Theory \( \text{PCT}^- + \text{INT} \)\(^{24}\) is the \( L_T \) theory extending PA with INT and the axiom
\[
\forall x \ \text{CC}(\langle x \rangle) \rightarrow \text{CC}(\langle x + 1 \rangle).
\]

Now, we have the following proposition:

PROPOSITION 4.9. The following theories are mutually \( \omega \)-interpretable:
1. WPT\(^-\) + INT
2. PCT\(^-\) + INT
3. ACA\(_0\).

Let us have a word of comment: the \( \omega \)-interpretation of PCT\(^-\) + INT in ACA\(_0\) is rather folklore and dates back to a paper of Mostowski [22] (however, in the context of GB and ZFC). The \( \omega \)-interpretation of ACA\(_0\) in WPT\(^-\) + INT is essentially the standard strategy of interpreting subsystems of second-order arithmetic in truth theories (which consists in, roughly speaking, identifying sets with their definitions; this strategy is used e.g., in [12]).

Proof of Proposition 4.9. We prove the interpretability of 1. in 2., then 2. in 3., and end with 3. in 1.

Interpretation of WPT\(^-\) + INT in PCT\(^-\) + INT.

Define
\[
T_{\text{WPT}^-}(x) = CC(\langle \text{dp}(x) + 1 \rangle) \land T(x).
\]
We claim that PCT\(^-\) + INT \( \vdash \) WPT\(^-\) + INT\[T_{\text{WPT}^-}/T\]. We show the axiom for disjunction and INT. The rest is very similar. We work in PCT\(^-\) + INT. Let us abbreviate
\[
\forall x (T_{\text{WPT}^-}(\phi(x)) \lor T_{\text{WPT}^-}(\neg \phi(x)))
\]
with tot\(_{\text{WPT}^-}(\phi(v))\). Fix sentences \( \phi, \psi \) and assume first that we have
\[
(T_{\text{WPT}^-}(\phi) \lor T_{\text{WPT}^-}(\psi)) \land \text{tot}_{\text{WPT}^-}(\phi) \land \text{tot}_{\text{WPT}^-}(\psi).
\]

\(^{23}\) The authors are grateful to an anonymous referee for the suggestions leading to this observation.
\(^{24}\) PCT is the acronym for “Progressive Compositional Truth”. 
Let \( d = \max\{\text{\(dp\)}(\phi), \text{\(dp\)}(\psi)\} \). In particular, we have \( \text{CC}(d + 1) \). Observe that by the progressiveness axiom of \( \text{PCT}^- + \text{INT} \) we have \( \text{CC}(d + 2) \) and since \( \text{dp}(\phi \lor \psi) = d + 1 \) and we have \( T(\phi) \lor T(\psi) \), we can conclude \( T(\phi \lor \psi) \) which gives us one implication from the axiom of disjunction.

Suppose now that we have \( \text{T}_{\text{WPT}^-}(\phi \lor \psi) \) and let \( d = \text{dp}(\phi \lor \psi) \). We have \( \text{CC}(d + 1) \). Observe that by the progressiveness axiom of \( \text{PCT}^- + \text{INT} \) we have \( \text{CC}(d + 2) \) and since \( \text{dp}(\phi \lor \psi) = d + 1 \) and we have \( T(\phi) \lor T(\psi) \), we can conclude \( T(\phi \lor \psi) \) which gives us one implication from the axiom of disjunction.

Let us now turn to \( \text{INT} \). Fix \( \phi(v) \) and assume \( \text{T}_{\text{WPT}^-}(\phi(0)) \land \forall x (\text{T}_{\text{WPT}^-}(\phi(x)) \rightarrow \text{T}_{\text{WPT}^-}(\phi(x + 1))) \). Let \( \text{dp}(\phi(v)) = d \). Since the depth of a formula is invariant under the substitution of terms for its free variables, we know that \( \text{CC}(d + 1) \) holds. By the internal induction axiom in \( \text{PCT}^- + \text{INT} \) we have \( \forall x T(\phi(x)) \), which concludes the proof.

Interpretation of \( \text{PCT}^- + \text{INT} \) in \( \text{ACA}_0 \).

This part is rather routine and we will only sketch it in order to assure that we can interpret the unrestricted axiom of induction in \( \text{ACA}_0 \). Define a formula \( T(X, x) \) with the second-order parameter \( X \)

\[
T(X, x) := \text{CC}(x)[y \in X / T(y)].
\]

In other words, \( T(X, x) \) asserts that \( X \) is the set of true formulae of complexity less than \( x \). Now define

\[
T_{\text{PCT}^-}(x) := \exists X \left( T(X, \text{dp}(x)) \land x \in X \right).
\]

Using arithmetical comprehension one can show that \( \text{ACA}_0 \models \text{PCT}^- [T_{\text{PCT}^-}(y) / T(y)] \).

It remains to show that \( \text{ACA}_0 \) proves the translation of \( \text{INT} \). The idea is very similar to the one used in the proof of the previous interpretation: fixing \( \phi(v) \) and assuming that

\[
T_{\text{PCT}^-}(\phi(x)) \land \forall x (T_{\text{PCT}^-}(\phi(x)) \rightarrow T_{\text{PCT}^-}(\phi(x + 1)))
\]

we conclude that there exists the set of true sentences of depth at most \( \text{dp}(\phi) + 1 \). Let us call it \( X \). Define the set \( Y \) using arithmetical comprehension:

\[
y \in Y \iff \phi(y) \in X.
\]

Therefore, by our assumption, we have

\[
0 \in Y \land \forall x (x \in Y \rightarrow x + 1 \in Y).
\]

The axiom of induction in \( \text{ACA}_0 \) gives us \( \forall x \ x \in Y \). This concludes the proof.

Interpretation of \( \text{ACA}_0 \) in \( \text{WPT}^- + \text{INT} \).

---

25 In fact, the axiom for disjunction is precisely where the argument for the interpretability of \( \text{PT}^- + \text{INT} \) in \( \text{ACA}_0 \) from [6] breaks down.

26 The results on model-theoretical strength of \( \text{PT}^- + \text{INT} \) and \( \text{PT}^- + \text{INT} \) \( \lvert_{\text{tot}} \) shows that the two axioms can differ in strength in some contexts.

27 See [6] and [23] for details.
We adopt the strategy from [12] and allow ourselves to rename variables in arithmetical sentences. Hence, if \( \tau \) denotes our translation, then we won’t obtain \( \phi^\tau = \phi \) for arithmetical sentences, but \( \phi^\tau \) will differ from \( \phi \) only with respect to the shape of bounded variables. Let \( L_2 \) denote the language of second-order arithmetic. Let us assume that

\[
X_0, X_1, \ldots
\]

are all the second-order variables of \( L_2 \) and

\[
x_0, x_1, \ldots
\]

are the first-order variables of this language. To simplify the argument we will use one designated first-order variable, \( x_0 \), and assume it is not the result of translating any variable of \( L_2 \). Let \( v \) be the code of \( x_0 \). We put:

\[
(X_i)^\tau = x_{2i+1} \\
(x_j)^\tau = x_{2j+2}.
\]

The translation we use is exactly as in [12], p. 107 with the exception that we interpret quantification over sets as quantification over total and consistent formulae, instead of arbitrary formulae. More precisely, we replace the condition:

\[
(\forall X_i \phi)^\tau = \forall x_{2i+1} (\text{Form}_{PA}^\leq (x_{2i+1}) \rightarrow \phi^\tau),
\]

with the following one:

\[
(\forall X_i \phi)^\tau = \forall x_{2i+1} (\text{Form}_{PA}^\leq (x_{2i+1}) \land \text{tot}(x_{2i+1}(v)) \land \text{cons}(x_{2i+1}(v)) \rightarrow \phi^\tau)
\]

where \( \text{cons}(x) \) is the natural dual of \( \text{tot}(x) \), i.e.,

\[
\text{cons}(\phi(v)) := \forall x \ 
eg (T(\phi(x)) \land T(\neg \phi(x))).
\]

Let us observe that the formula \( \text{Form}_{PA}^\leq (x) \land \text{tot}(x) \land \text{cons}(x) \) isolates precisely those formulae for which the predicate \( T \) commutes with the negation, i.e., we have

\[
\text{WPT}^- |\vdash \forall \phi(v) \ (\text{tot}(\phi(v)) \land \text{cons}(\phi(v)) \equiv \forall y (T(\neg \phi(y)) \equiv \neg T(\phi(y))).
\]

Under such a translation, Lemma 3.6, shows that \( \text{WPT}^- \) proves the translation of instances of the arithmetical comprehension scheme with semirelational formulae and this clearly suffices. Also, the translation of the induction axiom is simply the internal induction axiom restricted to total and consistent formulae. So it is provable even in a proper subtheory of \( \text{WPT}^- + \text{INT} \). We shall now show that it is model-theoretically conservative over \( \text{PA} \), hence can be used to model the use of the notion of truth in model theory as well.

\[\text{COROLLARY 4.10.} \ ACA_0 \ \text{is } \omega\text{-interpretable in } \text{WPT}^- + \text{INT}.\]

\[\text{Therefore, } \text{WPT}^- + \text{INT} \ \text{witnesses the expressive role of truth (at least no less than } \text{PT}^- + \text{INT} |\text{tot} \text{ does). We shall now show that it is model-theoretically conservative over } \text{PA}, \ \text{hence can be used to model the use of the notion of truth in model theory as well.}\]

\[\text{REMARK 4.11. In fact, all the rest theorems of this section can be easily derived from the mutual } \omega\text{-interpretability between } ACA_0 \text{ and } \text{WPT}^- + \text{INT}.\]
THEOREM 4.12. WPT$^-$ + INT is model-theoretically conservative over PA.

Proof. Let $\mathcal{M} \models PA$. For $b \in M$, let $b \in Tr'$ if and only if for some $t_0, \ldots, t_n$ such that

$$\mathcal{M} \models \bigwedge_{i \leq n} \text{CTerm}(t_i)$$

and some (standard!) $\phi(x_0, \ldots, x_n) \in \mathcal{L}_{PA}$

$$\mathcal{M} \models (b = [\phi(t_0, \ldots, t_n)]^T \land \phi(t_0^\circ, \ldots, t_n^\circ)).$$

Let us observe that with such a definition we have

$$(\mathcal{M}, Tr') \models \text{UTB}^-.$$  

To become the appropriate interpretation of WPT$^-$ truth predicate $Tr'$ requires only one small correction. Let $\sim_\alpha$ denote the arithmetical formula representing in PA the relation of two sentences being the same modulo renaming variables ($\alpha$-conversion). Let us define

$$b \in Tr$$

if and only if for some $\psi \in Tr'$, $\mathcal{M} \models b \sim_\alpha \psi$. Now, it can be easily shown that

$$(\mathcal{M}, Tr) \models \text{WPT}^- + \text{INT}.$$  

Indeed, compositional axioms are satisfied, since for every $x \in M$ such that $\mathcal{M} \models \text{Form}_{\leq 1} \mathcal{L}_{PA}(x)$

$$(\mathcal{M}, Tr) \models \text{tot}(x) \text{ if and only if for some } n \in \omega, \mathcal{M} \models \text{dp}(x) \leq n \quad (\ast)$$

and moreover $(\mathcal{M}, Tr) \models \text{UTB}^-$. Hence, in verifying compositional axioms we may use the fact that $\models$ is compositional. Let us check the axiom for $\lor$. Suppose $\phi = \psi \lor \theta$ and $(\mathcal{M}, Tr) \models T(\phi)$. Then there exists $\phi' \sim_\alpha \phi$ such that $T(\phi')$ and $\phi' = [\phi''(t_0, \ldots, t_n)]^T$ for some standard arithmetical formula $\phi''(x_0, \ldots, x_n)$ and $t_0, \ldots, t_n$ terms in the sense of $\mathcal{M}$. If so, then $\phi' = \psi' \lor \theta'$ such that $\psi' \sim_\alpha \psi'$ and $\theta \sim_\alpha \theta'$. Also, $\psi'$ and $\theta'$ are of the form $\psi''(t_0, \ldots, t_n)$ and $\theta''(t_0, \ldots, t_n)$, respectively. By UTB$, we have

$$\mathcal{M} \models \psi''(t_0^\circ, \ldots, t_n^\circ) \lor \theta''(t_0^\circ, \ldots, t_n^\circ).$$

Without loss of generality assume that $\mathcal{M} \models \psi''(t_0^\circ, \ldots, t_n^\circ)$. It means that $(\mathcal{M}, Tr) \models T(\psi)$ and consequently $(\mathcal{M}, Tr) \models T(\psi) \lor T(\theta)$. By $(\ast)$ we have

$$(\mathcal{M}, Tr) \models \text{tot}(\psi) \land \text{tot}(\theta) \land (T(\psi) \lor T(\theta)),$$

which completes the proof of one implication. Let us now assume that the above holds. Since we have $\text{tot}(\psi)$ and $\text{tot}(\theta)$ it follows that for some $n,k$,

$$\mathcal{M} \models \text{dp}(\psi) \leq n \land \text{dp}(\theta) \leq k.$$  

In particular, $\text{dp}(\phi) \leq \max\{n,k\} + 1$. Without loss of generality let us assume that $(\mathcal{M}, Tr') \models T(\psi)$. Let $\psi' \sim_\alpha \psi'$ and $\theta \sim_\alpha \theta'$ be standard arithmetical formulae such that $(\mathcal{M}, Tr') \models T(\psi')$. Reasoning as previously, we conclude that $(\mathcal{M}, Tr') \models T(\psi' \lor \theta')$ and hence

$$(\mathcal{M}, Tr) \models T(\psi \lor \theta)$$

which completes the proof of the compositional axiom for $\lor$. 
Let us now verify that $(\mathcal{M}, Tr) \models \text{INT}$. Fix an arbitrary formula $\phi(x)$ in the sense of $\mathcal{M}$ and assume that

$$(\mathcal{M}, Tr) \models T(\phi(0)) \land \forall x \ (T(\phi(x)) \to T(\phi(x + 1))).$$

It follows that $\mathcal{M} \models dp(\phi(x)) \leq n$ for some $n \in \omega$ and for some standard

$$\phi'_0(x, y_0, \ldots, y_k),$$

we have

$$\mathcal{M} \models \phi(0) \sim_{\alpha} \neg \phi'_0(0, t_0, \ldots, t_k) \uparrow$$

for some terms in $t_0, \ldots, t_k$ in the sense of $\mathcal{M}$. In particular, by UTB$^-$ we have

$$\mathcal{M} \models \phi'_0(0, t_0^\circ, \ldots, t_k^\circ).$$

We claim that also

$$\mathcal{M} \models \forall x \ (\phi'_0(x, t_0^\circ, \ldots, t_k^\circ) \to \phi'_0(x + 1, t_0^\circ, \ldots, t_k^\circ)).$$

Fix $x$ and assume that $\mathcal{M} \models \phi'_0(x, t_0^\circ, \ldots, t_k^\circ)$. It follows that

$$(\mathcal{M}, Tr) \models T(\phi(x)).$$

Hence, $(\mathcal{M}, Tr) \models T(\phi(x + 1))$. So for some $\phi'_1(x, y_0, \ldots, y_k)$ s.t.

$$\mathcal{M} \models \forall \phi'_1(x + 1, t_0, \ldots, t_k) \sim_{\alpha} \phi(x + 1, t_0, \ldots, t_k)$$

we have

$$\mathcal{M} \models \phi'_1(x + 1, t_0^\circ, \ldots, t_k^\circ).$$

But as satisfiability in a model is closed under $\alpha$-conversion, we get that

$$\mathcal{M} \models \phi'_0(x + 1, t_0^\circ, \ldots, t_k^\circ).$$

Hence, by induction in $\mathcal{M}$ we obtain

$$\mathcal{M} \models \forall x \ \phi'_0(x, t_0^\circ, \ldots, t_n^\circ)$$

which by UTB$^-$ again gives us $(\mathcal{M}, Tr') \models \forall x \ T(\phi'_0(x, t_0, \ldots, t_k))$. Hence also

$$(\mathcal{M}, Tr) \models \forall x \ T(\phi(x, t_0, \ldots, t_k)),$$

which ends the proof. \qed

In order to find a theory satisfying the Fischer-Horsten criterion we decided to switch the inner logic of the truth theory. It allowed to formulate a very natural theory of truth modelled after Weak Kleene Scheme. Is it possible to realize Fischer-Horsten desiderata using a compositional theory of truth extending PT$^-$? With the meaning we gave to the term “axiomatic theory of truth” we are not allowed to add more symbols to the language.\footnote{Without this restriction the answer is trivial: simply take PT$^-$ together with WPT$^- + \text{INT}$ but formulated with a different truth predicate symbol.}

For the moment, we leave it as an open problem.

§5. Acknowledgments. The research presented in this paper was supported by The National Science Centre, Poland (NCN), grant number 2014/13/B/HS1/02892.
**BIBLIOGRAPHY**

[1] Cantini, A. (1990). A theory of formal truth as strong as ID\textsubscript{1}. *The Journal of Symbolic Logic*, **55**(1), 244–259.

[2] Cieliski, C. (2015). The innocence of truth. *Dialectica*, **69**(1), 61–85.

[3] Cieśliński, C. (2015). Typed and untyped disquotational truth. In Achourioti, T., Galinon, H., Fernández, J. M., and Fujimoto, K., editors. *Unifying the Philosophy of Truth*. Dordrecht: Springer-Verlag, pp. 307–320.

[4] Cieśliński, C., Łełyk, M., & Wcisło, B. (2017). Models of PT\textsuperscript{−} with internal induction for total formulae. *The Review of Symbolic Logic*, **10**(1), 187–202.

[5] Feferman, S. (1991). Reflecting on incompleteness. *Journal of Symbolic Logic*, **56**(1), 1–49.

[6] Fischer, M. (2009). Minimal truth and interpretability. *Review of Symbolic Logic*, **2**, 799–815.

[7] Fischer, M. (2014). Truth and speed-up. *Review of Symbolic Logic*, **7**, 319–340.

[8] Fischer, M. (2015). Deflationism and instrumentalism. In Fujimoto, K., Fernández, J. M., Galinon, H., and Achourioti, T., editors. *Unifying the Philosophy of Truth*. Dordrecht: Springer, pp. 293–306.

[9] Fischer, M. & Horsten, L. (2015). The expressive power of truth. *Review of Symbolic Logic*, **8**, 345–369.

[10] Fitting, M. (1994). Kleene’s three valued logics and their children. *Fundamenta Informaticae*, **20**(1–3), 113–131.

[11] Fujimoto, K. (2010). Relative truth definability in axiomatic truth theories. *Bulletin of Symbolic Logic*, **16**, 305–344.

[12] Halbach, V. (2011). *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.

[13] Halbach, V. & Horsten, L. (2006). Axiomatizing kripke’s theory of truth. *Journal of Symbolic Logic*, **71**(2), 677–712.

[14] Halbach, V. & Leigh, G. E. (2014). Axiomatic theories of truth. In Zalta, E. N., editor. *The Stanford Encyclopedia of Philosophy* (Summer 2014 edition). Available at https://plato.stanford.edu/entries/truth-axiomatic/.

[15] Halbach, V. & Nicolai, C. (2018). On the costs of nonclassical logic. *Journal of Philosophical Logic*, **47**(2), 227–257.

[16] Horsten, L. (1995). The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In Cartois, P., editor. *The Many Problems of Realism (Studies in the General Philosophy of Science: Volume 3)*. Tilburg: Tilburg University Press, pp. 173–187.

[17] Kaufmann, M. (1977). A rather classless model. *Proceedings of American Mathematical Society*, **62**(2), 330–333.

[18] Kaye, R. (1991). *Models of Peano Arithmetic*. Oxford, UK: Clarendon Press.

[19] Ketland, J. (1999). Deflationism and tarski’s paradise. *Mind*, **108**, 69–94.

[20] Kripke, S. (1975). Outline of a theory of truth. *The Journal of Philosophy*, **72**(19), 690–716.

[21] Łełyk, M. & Wcisło, B. (2017). Models of weak theories of truth. *Archive for Mathematical Logic*, **56**(5), 1–26. doi:10.1007/s00153-017-0531-1.

[22] Mostowski, A. (1950). Some impredicative definitions in the axiomatic set theory. *Fundamenta Mathematicae*, **37**(1), 111–124.

[23] Schindler, R. ACA\textsubscript{0} and \Pi\textsubscript{1} \textsuperscript{1} – CA\textsubscript{0} and the semantics of arithmetic and BG and BG + \Sigma\textsubscript{1} \textsuperscript{1} – Ind, and the semantics of set theory. Unpublished manuscript. Available at www.math.uni-muenster.de/logik/Personen/rds/.
[24] Shapiro, S. (1998). Proof and truth: Through thick and thin. *Journal of Philosophy*, 95(10), 493–521.

[25] Shelah, S. (1978). Models with second order properties II. Trees with no undefined branches. *Annals of Mathematical Logic*, 14, 73–87.

[26] Smoryński, C. (1981). Recursively saturated nonstandard models of arithmetic. *The Journal of Symbolic Logic*, 46(2), 259–286.