ON EXPRESSING MODALITIES OVER ARITHMETIC

CARLO NICOLAI
LMU MUNICH

Abstract. The paper is concerned with the fine boundary between expressive power and reducibility of semantic and intensional notions in the context of arithmetical theories. I will consider three notions of reduction of a theory characterizing a semantic or a modal notion to the underlying base theory – relative interpretability, speed up, conservativeness – and highlight a series of cases where moving between equally satisfactory base theories and keeping the semantic or modal principles fixed yields incompatible results. I then consider the impact of the non-uniform behaviour of these reducibility relations on the philosophical significance we usually attribute to them.

1. Introduction

The claim that a reasonable arithmetical theory, in its twofold role of a mathematical and syntactic theory, cannot express satisfactory notions of truth, necessity, knowledge, belongs to the basic toolkit of the logically informed philosopher. Such a claim is certainly true under a specific reading of *expressing* in terms of explicit definitions: due to the well-known paradoxes of Tarski and Montague, there is no arithmetical (syntactic) formula satisfying the characterizing principles of these and closely related modal notions.

These limitative result impose that adequate theories of such notions result in *extensions* of the arithmetical base theory. If much is known about truth-theoretic extensions of a reasonable base theory,[1] much less investigated are extensions of a reasonable syntactic base with further intensional notions.[2] An initial study in this direction is the monograph [25], where it is considered the interaction of truth and predicates for necessity and possibility.

In this paper we are concerned with the fine boundary between expressive power and reducibility of intensional notions to the arithmetical or syntactic base theory. Over the years several authors have imposed adequacy conditions for the reducibility of intensional predicates to the base theory. For a recent example Fischer and

[1] The reader may consult [7] for a comprehensive overview. A proviso: first, the theories in [7] are all constructed over Peano Arithmetic PA and it is not always immediate to extend the results there to other base theories.

[2] Note on terminology: In what follows, we employ the adjective 'intensional' and 'modal' in an inclusive sense, familiar to medieval authors such as Ockham, which comprehend also truth and propositional attitudes.

Horsten, in [4], consider a theory of truth to be adequate if it is semantically conservative over the base theory – i.e. any model of the base theory can be expanded to a model of the theory of truth – and at the same time expressively irreducible to it in the sense of being not relatively interpretable in it and having additional emerging expressive features such as non-elementary speed-up over it.[3]

In the following sections, we will show that there are non-uniformity phenomena concerning the combinations of these notions of reduction, i.e. conservativeness, relative interpretability, non-elementary speed-up, that cast serious doubts on the very possibility of analyzing intensional notions in terms of their reducibility to the base theory. In a nutshell, we will mainly highlight the fact that by moving from a reasonable base theory to another one may unintentionally go from a reducible to an irreducible intensional notion. If the adequacy of the truth or modal theory is evaluated in terms of its reducibility, as Horsten and Fischer propose, the variation of seemingly innocuous assumptions on the bearers of semantic and modal ascriptions may result in an incoherent body of information. This may seriously compromise any attempt of extracting philosophical information from modal and semantic theories. In the concluding section, after presenting the situation more clearly, we will suggest a possible way out of this riddle.

Note to the reader. To highlight the main conceptual points and enhance readability, I have tried to reduce formal details to the bare bones. I have nonetheless kept the main points of the proofs of the results I employ by embedding them in the body of the text. The reader interested in full proofs may refer to the bibliography accompanying the informal arguments. Also, the reader may find certain conceptual moves in §2-4 rather hasty. This is because I have decided to condense technicalities and the necessary preliminaries for a proper discussion in these sections and then reconstruct the main line of reasoning, with more care on the philosophical points, in the final section.

2. Bearers of modal ascriptions and the three case studies

In what follows, we will consider base theories B that are sufficiently strong to develop a nice theory of the bearers of truth and modal ascriptions. Such theories are generally required to articulate the properties of expression types – that will be in what follows the objects to which semantic and modal notions apply – via suitable coding. This, in turn, requires the possibility of coding finite sequences of objects and, therefore, a modicum of number theory to represent functions such as projections. A useful notion capturing in one go all these desiderata is the notion of *sequential* theory [20, 27], which is formally defined as a theory that relatively

[3] Further references abound: we go from the well-known conservativeness argument discussed in [23, 13, 3] to more recent discussions of deflationism [9, 18].

interprets *Adjunctive Set Theory*

$$\text{(AS1)} \qquad \exists y \forall u (u \notin y)$$

$$\text{(AS2)} \qquad \exists v \forall u (u \in v \leftrightarrow u \in x \vee u = y)$$

without relativizing quantifiers and preserving identity. In what follows we will therefore stipulate that all base theories that we employ are sequential.

A particularly nice feature of sequential theories is that they can formalize, via a suitable interpretation, the usual development of the syntax of first-order theories as it is required, for instance, for Gödel's second incompleteness theorem. To see this, it suffices to notice that AS interprets theories such as Buss' S^1_2 or $\mathsf{I}\Delta_0 + \Omega_1$, which are precisely up to the task of formalizing syntactic notions and operations. This interpretation guarantees the right syntactic structure to the objects of truth and modalities.[4] This fact, combined with the possibility of coding sequences of all objects in the domain of a sequential theory, makes it possible to formulate the usual clauses for satisfaction that are a necessary condition for extending our base sequential theory with semantic or modal predicates.

Inside the family of sequential theories, we will distinguish between schematically presented or *schematic* theories featuring a finite set of axioms together with a finite set of axiom schemata, and *finitely axiomatized* theories. More precisely, we will mostly deal with a subclass of schematic theories that we will call *inductive*, in the sense that they satisfy (possibly via translation) the full induction schema for $\mathcal{L}_B$:

$$(1) \qquad \varphi(0) \wedge \forall x \in \omega\, (\varphi(x) \leftrightarrow \varphi(x+1)) \rightarrow \forall x \in \omega\, \varphi(x)$$

The possibility of a translation suggests, of course, that the set of natural numbers ω does not necessarily have to be available in B; in fact, on many occasions, it won't be. At any rate, inductive (sequential) base theories are capable of developing partial truth predicates satisfying Tarski's inductive clauses for each of their finite subtheories B_0. This, in turn, entails that they can prove a canonical consistency statement for B_0. By the second incompleteness theorem, therefore, the class of inductive and the class of finite sequential theories are disjoint. Examples of inductive base theories are Peano arithmetic PA, Zermelo-Fraenkel set theory ZF, full second-oder arithmetic Z_2; examples of finitely axiomatized sequential theories are elementary arithmetic EA, the extension of PA with arithmetical comprehension ACA_0, and Von Neumann-Gödel-Bernays set theory NBG.

We assume a fixed Gödel coding and a formalization of the syntax for our base theory B as it can be standardly carried out in S^1_2. In particular, we write $\ulcorner e \urcorner$ for the term of $\mathcal{L}_B$ representing the Gödel code of the $\mathcal{L}_B$-expression e. We assume that $\mathcal{L}_B$ contains, besides its characterizing arithmetical function symbols, also finitely many additional function symbols for syntactic operations: for instance, $\dot{\neg}$ is a function symbol expressing the syntactic operation of negation. It is also convenient to have

[4] For details on the subexponential theories S^1_2, $\mathsf{I}\Delta_0 + \Omega_1$ the standard reference is still [11, §V].

in our language a function symbol for the operation of replacing a term z for a free variable v in a formula φ.

We will mostly deal with extensions of B with principles governing semantic and modal notions. For simplicity, I will refer to these extensions as *modal theories*. For our purposes it suffice to consider the language $\mathcal{L}_B \cup \{\Box\}$, where $\Box$ is a fresh unary predicate. We abbreviate $\Box\ulcorner\varphi\urcorner$ with $\Box\varphi$ and, similarly $\Box\ulcorner\varphi(\dot{x})\urcorner$ – that is the result of formally substituting in $\varphi(v)$ the variable v with the numeral for x, usually expressed via the substitution function and the numeral function $\mathsf{sub}(\ulcorner\varphi(v)\urcorner, \mathsf{num}(x))$ – with $\Box\varphi(x)$.

With these little preliminaries at hand, we can introduce the extensions of our base theories we will be mainly interested in. The first set of principles that we consider is encompassed in the schema:

$$(\mathsf{M}) \qquad \forall x(\Box\varphi(x) \leftrightarrow \varphi(x))$$

for $\varphi \in \mathcal{L}_B$. Several theories of truth, but also some of the theories of necessity in [25], feature this principle. (M) is manifestly plausible for alethic modalities such as truth and necessity, it essentially tells us that the extension of $\Box$ restricted to standard sentences of $\mathcal{L}_B$ agrees with the $\mathcal{L}_B$-truths of the modal theory. We will call $\mathsf{M}[B]$ the result of adding (M) to B.

The second group of principles we are interested in reflects the possibility of having a uniform distribution of the semantic or modal predicate over the logical connectives starting with with truths of B and building up inductively the extension of $\Box$. In this case, we treat negation on a case by case manner:

(G1) $\forall x\,\big((\Box R(x) \leftrightarrow R(x)) \wedge (\Box\neg R(x) \leftrightarrow \neg R(x))\big)$ for any relation $R \in \mathcal{L}_B$

(G2) $\forall x, y\,\big(\mathsf{Sent}_{\mathcal{L}_B}(x \underset{\cdot}{\wedge} y) \to (\Box(x \underset{\cdot}{\wedge} y) \leftrightarrow \Box x \wedge \Box y)\big)$

(G3) $\forall x, y\,\big(\mathsf{Sent}_{\mathcal{L}_B}(x \underset{\cdot}{\wedge} y) \to (\Box\underset{\cdot}{\neg}(x \underset{\cdot}{\wedge} y) \leftrightarrow \Box\underset{\cdot}{\neg}x \vee \Box\underset{\cdot}{\neg}y)\big)$

(G4) $\forall v, x\,(\mathsf{Sent}_{\mathcal{L}_B}(\underset{\cdot}{\forall}vx) \to (\Box(\underset{\cdot}{\forall}vx) \leftrightarrow \forall y\ \Box\, x(y/v)))$

(G5) $\forall v, x\,(\mathsf{Sent}_{\mathcal{L}_B}(\underset{\cdot}{\forall}vx) \to (\Box(\underset{\cdot}{\neg}\underset{\cdot}{\forall}vx) \leftrightarrow \exists y\ \Box\, \underset{\cdot}{\neg}x(y/v)))$

(G6) $\forall x\,\big(\mathsf{Sent}_{\mathcal{L}_B}(x) \to (\Box\underset{\cdot}{\neg}\underset{\cdot}{\neg}x \leftrightarrow \Box x)\big)$

Since the predicate $\Box$ occurs only positively in the clauses G1-G5, the resulting cluster of theories $\mathsf{G}[B]$, although only dealing with typed predicates, may be taken to capture the *grounded* development of $\Box$. We will discuss shortly the choice of a typed predicate, that is a predicate that only applies to $\mathcal{L}_B$ sentences and not to sentence containing $\Box$.

Finally, we will consider the theories $\mathsf{C}[B]$ obtained by extending B with the axioms G1, G2, G4 and the axioms stipulating the full commutativity of $\Box$ with negation:

$$(\neg) \qquad \forall x(\mathsf{Sent}_{\mathcal{L}_B}(x) \to (\Box\underset{\cdot}{\neg}x \leftrightarrow \neg\,\Box\, x))$$

It is clear that, in $\mathsf{C}[B]$, $\Box$ is not treated positively. However, it will be treated *compositionally*. This property is clearly desirable for truth and arguably for other alethic modalities such as necessity.

The theories $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[B]$ will be the three case studies we will be occupied with in the rest of the paper. Two remarks are in order. On the one hand, since we will vary the base theory B while keeping the principles for $\Box$ fixed, we need to be clear about the role of the nonlogical axiom schemata of B, if present. Unless otherwise specified, we do *not* extend nonlogical axiom schemata of B to $\Box$. On the other, as it is clear from the principles above, we will impose a restriction to the applicability of the predicate $\Box$ in the theories $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[B]$: in particular, these theories will force no sentence containing $\Box$ into the extension of $\Box$ itself. In other words, the theories $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[B]$ are *typed* treatments of $\Box$. Both the non extension of nonlogical axiom schemata of B to $\Box$ and the typed nature of our theories are motivated by our interest in the thin line separating reducible and non-reducible modal theories (to the base theory B): extending schemata will in fact result in non-reducible theories, whereas considering type-free notions will often only complicate the study of the reductions we are interested in without affecting the overall conceptual point, since the theories $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[B]$ are obviously contained in their type-free versions and in virtually all known semantic and modal theories.

In the next three sections we consider three well-known notions of reduction of the modal theory to the base theory B: conservativeness, relative interpretability, non elementary speed-up.

3. Deductive strength

The proof-theoretic notion of conservativeness provides one with a precise sense in which a modal extension of a reasonable base theory B may involve *insubstantial* concepts, namely concepts that do not play a significant role in the explanation of non-modal facts. This reading is reminiscent of certain formal renderings of truth-theoretic deflationism (see for instance Horwich's [10] and [23, 13, 9]) according to which the truth predicate should not play a substantial role in the explanation of non-semantic facts. A modal theory T is *conservative* over B if any theorem φ in the language of B that is provable in T is already provable in B alone. Of course there is no consensus on whether notions such as truth, necessity, possibility and the like ought to be insubstantial in the sense just hinted at. For our concerns, in fact, it suffices that such interpretations exist.

For B sequential, the theory $\mathsf{M}[B]$ is a conservative extension of B. This can be established by a well-known argument dating back to Tarski's [26]. In a proof of a sentence φ of $\mathcal{L}_B$, in fact, only finitely many occurrences of the schema (M) can occur. This means that there is a finite set of formulas $\varphi_1, \dots, \varphi_n$ that occur in instances of (M). Then one can simply define in B a predicate

$$\text{(2)} \qquad \mathsf{P}(x,y) :\leftrightarrow (x = \ulcorner \varphi_1(\dot{y}) \urcorner \wedge \varphi_1(y)) \vee \dots \vee (x = \ulcorner \varphi_n(\dot{y}) \urcorner \wedge \varphi_n(y))$$

and replace occurrences of $\Box\mathsf{sub}(\ulcorner\varphi_i\urcorner, \mathsf{num}(y))$ in the proof of φ with $\mathsf{P}(\ulcorner\varphi_i\urcorner, y)$. The resulting proof is a legitimate proof of φ in B, witnessing the conservativity of $\mathsf{M}[B]$ over B.

Also the conservativity of $\mathsf{G}[B]$ over sequential B can be easily established, although via standard semantic considerations. By slightly extending an argument contained in [1], in fact, any model of $\mathcal{M} \models B$ can be expanded to a model of $(\mathcal{M}, S)$ of $\mathsf{G}[M]$, where S is the extension of the predicate $\Box$. The fundamental reason for this is that a set of $\mathcal{L}_B$ sentences satisfying the $\mathsf{G}[M]$ axioms can be characterized as a fixed point of a positive inductive definition, where $f^{\mathcal{M}}$ stands for the denotation in M of the function symbol f of $\mathcal{L}_B$ and we assume an expanded language $\mathcal{L}_B^+$ featuring names for all objects in the domain $|\mathcal{M}|$ of $\mathcal{M}$:

$$
\begin{aligned}
a \in X \Leftrightarrow\ & a \text{ is a sentence of } \mathcal{L}_B^+, \text{ and}\\
& \Big(a \text{ is a true atomic formula or negated atomic formula, or}\\
& a \text{ is } b \underset{\cdot}{\wedge}^{\mathcal{M}} c \text{ and } b \in X \text{ and } c \in X \text{ for some sentences } b, c, \text{ or}\\
& a \text{ is } \underset{\cdot}{\neg}^{\mathcal{M}}(b \underset{\cdot}{\wedge}^{\mathcal{M}} c) \text{ and } \underset{\cdot}{\neg}^{\mathcal{M}} b \in X \text{ or } \underset{\cdot}{\neg} c^{\mathcal{M}} \in X \text{ for some sentences } b, c, \text{ or}\\
& a \text{ is } \underset{\cdot}{\forall} bc \text{ and for all } d \in |\mathcal{M}|, \mathsf{sub}^{\mathcal{M}}(c, \mathsf{num}^{\mathcal{M}}(d)) \in X\\
& \ \text{for } c \text{ a formula with one free variable and } b \text{ a variable, or}\\
& a \text{ is } \underset{\cdot}{\neg}^{\mathcal{M}} \underset{\cdot}{\forall} bc \text{ and for some } d \in |\mathcal{M}|, \underset{\cdot}{\neg}^{\mathcal{M}}\mathsf{sub}^{\mathcal{M}}(c, \mathsf{num}^{\mathcal{M}}(d)) \in X\\
& \text{for } c \text{ a formula with one free variable and } b \text{ a variable}\Big)
\end{aligned}
$$

In a fixed point of this inductive definition, that is a set S such that $(\mathcal{M}, S)$ satisfies the above equivalence, we will have that, for instance, a sentence $\varphi \wedge \psi$ is in S if and only if φ is in S and ψ is in S, and similarly for all other axioms of $\mathsf{G}[B]$.

The conservativeness of the theory $\mathsf{C}[B]$ over suitable B cannot be achieved so easily. For a substantial chunk of sequential theories, that is the ones extending (either directly or via interpretation), a specific subsystem of first-order arithmetic called $\mathsf{I}\Delta_0(\mathsf{exp})$ or EA, Leigh [14] has produced an argument showing how to eliminate cuts on formulas of the form $\Box\varphi$ in proofs of $\mathcal{L}_B$-formulas. The role of EA, a system designed to capture exactly a proper subclass of primitive recursive functions, the so-called Kalmar's elementary functions, is roughly the one of controlling the complexity of boxed formulas in derivations so that the usual induction on the length of the derivation of $\mathcal{L}_B$-theorems to push $\Box$-cuts upwards could be carried out.[5] It is still unknown whether the theory $\mathsf{M}[B]$ is conservative over B for *all* sequential B.[6]

[5] It should be noticed that Halbach's [8] contains a cut-elimination argument that was later shown to contain a gap. The problem there was exactly this absence of a suitable machinery to control the complexity of boxed formulas in proofs of $\mathcal{L}_B$-formulas.

[6] For the reader not familiar with subsystems of first-order arithmetic, it may be useful to know that it's easy to find theories that are proper sub-theories of EA but still sequential. One example is Buss' theory S^1_2 mentioned above.

MODAL THEORY OVER A REASONABLE B	INDUCTIVE	FINITE
$\mathsf{M}[B]$	✓	✓
$\mathsf{G}[B]$	✓	✓
$\mathsf{C}[B]$ $(\supseteq \mathsf{I}\Delta_0(\mathsf{exp}))$	✓	✓

Table 1. Conservativeness.

Table 3 summarizes the results sketched in the last few paragraphs. My aim in this paper is to highlight how conservativeness and other notions of proof-theoretic reduction are highly sensitive to the choice of the underlying, sequential base theory. Conservativeness, however, appears to behave in a very stable way. The overall analysis that we are going to propose in the concluding paragraph will be that this stability is only due to the coarse-grained nature, if compared with other notions, of the relation of conservativeness.

If we move to a setting in which schemata of B may be extended to $\Box$, however, non-uniformity phenomena appear also in the context of the conservativeness of the modal theories considered over B. We consider one simple example: the theory $\mathsf{C}[\mathsf{PA}]^+$, for instance, where the $^+$ denotes the extension of the induction schema of PA to the $\Box$, is strong enough to formalize the soundness of PA by reading $\Box$ as truth. Therefore, $\mathsf{C}[\mathsf{PA}]^+$ will prove $\mathsf{Con}(\mathsf{PA})$ and will not be conservative over PA.[7] However, if we leave arithmetic and move to base theories that are able to prove strong forms of reflection, such as ZF or Z_2, the situation changes. It is well-known, in fact, that ZF proves, for any sentence φ of the language of set theory, its equivalence to its relativization φ^{V_α} for some limit ordinal α (see for instance [12, Thm. 12.14]); similarly, in Z_2 or, equivalently, Π^1_ω-CA, we can prove that for any sentence φ of the language of second-order arithmetic has a countable ω-model (see [24, Lem. VIII.5.2]).

Now consider, seeking a contradiction, a sentence φ of the language $\mathcal{L}_2$ of Z_2 such that, for instance, $\mathsf{C}[\mathsf{Z}_2]^+$ proves φ and Z_2 does not prove φ. It's important to notice that $\mathsf{C}[\mathsf{Z}_2]^+$ features the schema of $\mathcal{L}_2$-induction extended to the $\Box$ but *not* the extended full-comprehension schema. Let A be the finite subsystem of $\mathsf{C}[\mathsf{Z}_2]^+$ proving φ — which is given by the compactness theorem — and the set B of the axioms of Z_2 employed in the proof of φ. By the ω-reflection principle Z_2 proves that there is an ω-model of B. The truth predicate associated with this model suffices to interpret A in Z_2, including the instances of the extended induction. Therefore, Z_2 proves φ after all. The argument transfers to ZF without essential modifications except for the

[7] The argument is folklore, see for instance [7, §8].

use of the Levy-Montague reflection principle instead of the ω-reflection principle: again we highlight that $\mathsf{C}[\mathsf{ZF}]^+$ is formulated by means of a schema of ω-induction extended to $\Box$ but without extending the schemata of separation and replacement of ZF.

It should be now clear how the results mentioned in the last paragraphs amount to a case of non-uniformity for the notion of conservativeness relative to the choice of base theories, at least in relation to one of our modal theories: if $\mathsf{C}[\mathsf{PA}]^+$ is non conservative over PA, the theories $\mathsf{C}[\mathsf{ZF}]^+$ and $\mathsf{C}[\mathsf{Z}_2]^+$ are indeed conservative over their respective base theories. One objection is in order: how can one justify the extension of the *ω-induction* schema of, say, ZF to $\Box$ and the simultaneous non-extension of the other non-logical schemata of ZF? One straightforward answer may be that, since $\Box$ has to be interpreted as truth, necessity, possibility, knowledge, and the like, its range of application are *syntactic* objects: as a consequence, since it is well-known that the structure of syntactic objects — strings of a finite alphabet in particular — is isomorphic to that of natural numbers, the extended principle of induction can be easily justified as a syntactic principle (see also [6] on this point). By contrast, surely we would consider, say, the full replacement schema a syntactic principle.

4. Conceptual reducibility

The connection between the conceptual reducibility of a modal notion to the arithmetical resources and the *relative interpretability* of a modal theory to the base theory has been recently highlighted by a number of authors (see for instance [9, 4, 18]). Informally, a theory U is relatively interpretable in a theory V if there is a translation of the primitive concepts of the language of U into the language of V that commutes with propositional connectives, possibly relativizes quantifiers, and preserves theoremhood.[8] U is locally interpretable in V if every finite subtheory of U is interpretable in V.

As we just did for conservativeness, we ask ourselves whether $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[B]$ are interpretable in B. To do so, however, we need to introduce a bit of terminology. A formula $\varphi(v)$ is said to be *progressive* in U if U proves $\varphi(0)$ and $\forall x\,(\varphi(x) \to \varphi(x+1))$; $\varphi(v)$ is a *cut* in U if it is progressive in U and downwards closed under $\leq$, that is if U proves $\forall y, x(y \leq x \wedge \varphi(x) \to \varphi(y))$. A remarkable fact due to Robert Solovay is that, in extensions V of Robinson's arithmetic Q, for every every inductive formula $\varphi(v)$ one can find a V-cut ψ such that V proves $\forall x(\psi(x) \to \varphi(x))$ (see [11, Lem. 5.9]).

It turns out that, when relative interpretability is involved, the distinction between inductive and finitely axiomatized sequential theories matters. If B is inductive, in fact, it can prove the consistency of any of its finite subtheories. By a well-known result often called Orey's compactness theorem, if U is locally interpretable

[8] For a formally precise definition, see for instance [27].

in V and V is inductive, then U is also relatively interpretable in it.[9] Therefore, if B is inductive, the proof of the conservativeness of $\mathsf{M}[B]$ over B, which can be easily seen to amount to a proof of local interpretability, is indeed a proof of the interpretability of $\mathsf{M}[B]$ in B. If, by contrast, B is finite, a recent unpublished argument by Albert Visser shows that $\mathsf{M}[B]$ is *not* interpretable in it.[10] For B finite, in fact, $\mathsf{M}[B]$ can define a cut $\mathcal{I}$ that is the intersection of all B-definable cuts. At the same time, by a result of Pudlák [21], for *each* $n \in \omega$, B can define a cut $\mathcal{J}$ such that $B \vdash \mathsf{Con}^{\mathcal{J}}_n(B)$ – where $\mathsf{Con}^{\mathcal{J}}_n(B)$ says that there is no proof in $\mathcal{J}$ of contradiction from axioms of Gödel number smaller than n and with a proof whose elements contain at most n quantifiers. The two claims just made entail that $\mathsf{M}[B]$ can define the intersection of all such $\mathcal{J}'s$. Therefore B can interpret, by relativizing all quantifiers to the cut $\mathcal{I}$, the theory $\mathsf{S}^1_2 + \{n \in \omega \mid \mathsf{Con}_n(B)\}$. But, by another result of Pudlák (see [21, Cor. 3.5]), no finitely axiomatized sequential theory T can interpret $\mathsf{S}^1_2 + \{n \in \omega \mid \mathsf{Con}_n(T)\}$; so if $\mathsf{M}[B]$ was interpretable in B, we would get a contradiction by the transitivity of interpretability.

We have therefore just seen a first non-uniformity result: if B is inductive, $\mathsf{M}[B]$ is interpretable in B. If it is finite, $\mathsf{M}[B]$ is not interpretable in B. We find a similar scenario when we move to the theories $\mathsf{G}[B]$ and $\mathsf{C}[B]$. For our purposes it is better to start with considering $\mathsf{C}[B]$.

If B is inductive, it will prove the consistency of any of its finite subtheories. By Leigh's proof of the conservativity of $\mathsf{C}[B]$ over B considered above, therefore, which is formalizable in EA, we have for B containing EA,

$$B \vdash \mathsf{Con}(B_0) \rightarrow \mathsf{Con}(\mathsf{C}[B_0]) \tag{3}$$

where B_0 is a finite subsystem of B. Therefore $\mathsf{C}[B]$ will prove $\mathsf{Con}(\mathsf{C}[B_0])$ and so it will also be reflexive. It is well-known, however, that the relative interpretability of T_0 in T_1, with T_0, T_1 reflexive, follows from the arithmetical Π_1-conservativity of T_0 over T_1. This therefore yields the interpretability of $\mathsf{C}[B]$ and $\mathsf{G}[B]$ in B for B inductive.

By contrast, if B is a finitely axiomatized sequential theory, $\mathsf{C}[B]$ and $\mathsf{G}[B]$ are not interpretable in B. We first consider the idea of the argument for $\mathsf{C}[B]$. By the inclusion of $\mathsf{M}[B]$ in $\mathsf{C}[B]$, the axioms of B are all in the extension of $\Box$. Assuming that B is formulated in a calculus in which Modus Ponens is the only rule of inference, we move to a $\mathsf{C}[B]$-definable cut $\mathcal{K}$ in which all (codes) of logical axioms of B are in the extension of $\Box$. We close $\mathcal{K}$ under provability in $\mathsf{C}[B]$, by moving to another cut $\mathcal{N} \subset \mathcal{K}$: this is possible by Solovay's method because the property 'the

[9] Here's a proof: Let's assume that U is locally interpretable in V. In V, either $\mathsf{Con}(U)$ or $\neg\mathsf{Con}(U)$. If the former, an interpretation can be found via the Henkin-Feferman arithmetized completeness theorem. If the latter, then there is a finite $V_0 \subseteq V$ such that $V \vdash \mathsf{Con}(V_0) \rightarrow \mathsf{Con}(U_0)$ for all finite $U_0 \subset U$. Therefore V proves $\mathsf{Con}(U_0)$. But then, by the well-known argument due to Feferman [2], one can find an intensional consistency statement $\mathsf{Con}^*(U)$ such that $V \vdash \mathsf{Con}^*(U)$. We can then employ the Henkin-Feferman construction again.

[10] Personal communication.

last element of a B-proof x in $\mathcal{K}$ is in the extension of $\Box$' is provably progressive in $\mathsf{C}[B]$. All theorems of $\mathsf{C}[B]$ that belong to $\mathcal{N}$, therefore, are in the extension of $\Box$. If a proof of a falsity $\bot$ was in $\mathcal{N}$, then, $\Box\bot$ would be provable in $\mathsf{C}[B]$ and so would $\bot$. In other words, we can prove the consistency of B relative to $\mathcal{N}$ in $\mathsf{C}[B]$. By Pudlák's result, therefore, B cannot interpret $\mathsf{C}[B]$.[11]

Let's now turn to $\mathsf{G}[B]$ for finitely axiomatized B: it also contains $\mathsf{M}[B]$, and so it proves that all nonlogical axioms of B are in the extension of $\Box$. To replicate the argument just given for the non-intepretability of $\mathsf{C}[B]$ in B, we would need a workable axiom of full commutation of $\Box$ with negation.[12] Since we cannot have it in full, we notice that the 'formula'

'for every sentence φ of $\mathcal{L}_B$ of logical complexity $\leq x$: $\neg \Box \varphi$ if and only if $\Box\neg\varphi$'

is provably progressive in $\mathsf{G}[B]$. Therefore we can find a $\mathsf{G}[B]$-cut $\mathcal{H}$ such that the formulas belonging to it enjoy full commutation of negation. We intersect this cut with the analogue of the cut $\mathcal{K}$ capturing a portion of the true logical axioms of B. We can then close $\mathcal{H} \cap \mathcal{K}$ under provability and prove the consistency of B on a cut in $\mathsf{G}[B]$ as well, which yields its non-interpretability in B (for more details we refer to [16]).

Relative interpretability, that we have associated with the conceptual reducibility of the modal notion to the syntactic resources of the underlying base theory, displays therefore a deeply non-uniform behaviour. Table 4 summarizes the situation: For B inductive and 'reasonable' — that is interpreting EA — $\mathsf{M}[B]$, $\mathsf{C}[B]$ and $\mathsf{G}[B]$ are all interpretable in B. For all sequential and finitely axiomatized B, $\mathsf{M}[B]$, $\mathsf{G}[B]$ and $\mathsf{C}[B]$ are not interpretable in B.

The message to extract from these results is clear: if one associates the conceptual reducibility of a modal notion characterized via a modal theory to the relative interpretability of the latter into its base theory, as [9] and [4] suggest, one has to face the fact that the *very same* principles for $\Box$ may be associated to a conceptually reducible or irreducible notion depending on facts that have nothing to do with the characterization of $\Box$. Inductive and finitely axiomatized sequential theories, for all we know, satisfy the desiderata imposed to the characterization of the bearers of modal ascriptions that can be found in the literature (see again [9, 7, 25]). We will see in the concluding section that perhaps something is indeed missing in these accounts.

5. Instrumentalism

If conservativeness behaves uniformly, relative interpretability exhibits a discouraging discontinuity. As a possible tie-breaker, I consider the capability of the modal

[11] For a full argument, see [19].

[12] Notice for instance, that a form of full commutation with negation is needed even for the principle

$$(\mathsf{K}) \qquad \forall\varphi,\psi(\Box(\varphi\to\psi)\wedge\Box\varphi\to\Box\psi)$$

that is required to close provability under $\Box$.

MODAL THEORY OVER A REASONABLE B	INDUCTIVE	FINITE
$\mathsf{M}[B]$	✓	✗
$\mathsf{G}[B]$	✓	✗
$\mathsf{C}[B]$	✓	✗

TABLE 2. Relative Interpretability.

theories of shortening proofs of theorems of B, the so-called *speed-up* phenomenon. As Fischer suggests in [5] in relation to truth, a notion that possesses this property would enable us to significantly enhance our expressive resources becoming an indispensable *instrument* to express in a more concise and tractable way some facts concerning an underlying ontology: in our case the structure of the bearers of modal ascriptions. We will soon consider some concrete examples. The leading question of this section is, therefore: when moving from an inductive to a finite base theory B, is the capability of $\mathsf{M}[B]$, $\mathsf{G}[B]$, and $\mathsf{C}[M]$ of shortening B-proofs affected?

As before, we briefly introduce some technical tools and terminology from [22, 5]. Given a Hilbert-style formulation of a theory U, we let $|\varphi|_U$ be the shortest (code of a) U-proof of φ, if that proof exists at all. Given theories U, V with $U \subseteq V$, we say that V has *at most polynomial speed up* over U if there is a polynomial p such that, for φ provable in U, if $|\varphi|_V \leq n$, then $|\varphi|_U \leq p(n)$. We are encouraged from complexity theory to set aside polynomial speed up. By contrast, *non-elementary speed up* is not to be overlooked: V has non-elementary speed up if the are $\mathcal{L}_U$-formulas $\varphi_1, \dots, \varphi_n$ and no function F of elementary growth rate such that, if $|\varphi|_V \leq n$, then $|\varphi|_U < F(n)$. By elementary growth rate, I mean a function $f(x)$ that can be majorized by a superexponential function 2^x_m.[13]

As in the case of conservativeness, the main question of this section can be addressed uniformly for $\mathsf{M}[B]$ for sequential B, regardless of their being finite or inductive. A version of the argument witnessing the conservativeness of $\mathsf{M}[B]$ over B sketched on page 5 can be given so that when we transform a $\mathsf{M}[B]$ proof, say of length n, to a B-proof, the size of the latter only grows by a polynomial in n. The essential ingredient of the argument is the observation that for any formula $\varphi(\vec{x})$ of $\mathcal{L}_B$ we can find a partial truth predicate for it such that the proof of its 'disquotational' property

$$\mathsf{Tr}_\varphi(\varphi(x)) \leftrightarrow \varphi(x)$$

can be proved in B with a proof of size polynomial in the code of φ (see [22, Thm. 3.3.1]). Therefore, for B sequential, $\mathsf{M}[B]$ *has at most polynomial speed up over B*.

[13] Where $2^x_0 = x$ and $2^x_{y+1} = 2^{2^x_y}$.

MODAL THEORY OVER A REASONABLE B	INDUCTIVE	FINITE
$\mathsf{M}[B]$	✗	✗
$\mathsf{G}[B]$	?✗?	✓
$\mathsf{C}[B]$	?✗?	✓

TABLE 3. Non-elementary speed up over B.

As before, the situation for $\mathsf{G}[B]$ and $\mathsf{C}[B]$ is more complex. When B is finite and sequential, we have seen that they prove the consistency of B on a cut definable in $\mathsf{G}[B]$ or $\mathsf{C}[B]$. By contrast, by a fundamental result of Friedman and Pudlák, the consistency statements $\mathsf{Con}_{2_n}(B)$ are, for each n and some constant c, only provable in B with a proof of size greater than 2_n^c (see [22, Thm. 6.2.3]). However, for theories that prove the consistency of B on a cut, proofs of $\mathsf{Con}_{2_n}(B)$ become *linear* in n.[14] Therefore $\mathsf{C}[B]$ and $\mathsf{M}[B]$ have non-elementary speed up over B for B finitely axiomatized and sequential.

What happens when B is inductive? Unfortunately we don't have a clear answer, but only the strong conjecture that the proofs of the interpretability of $\mathsf{C}[B]$ in B, for B inductive, may give a polynomially bounded reduction of $\mathsf{C}[B]$-proofs to B-proofs. This would in turn yield the lack of non-elementary speed up of $\mathsf{G}[B]$ over B for inductive B.

Table 5 summarizes the situation for speed up, which is more open than the ones for relative interpretability and conservativeness, although the proofs of the interpretability of $\mathsf{C}[B]$ in B for B inductive strongly support that idea that also speed up has a non-uniform behaviour. From the conceptual point of view, again if our conjectures are true, we would have that the capability of the predicate $\Box$ of shortening proofs and therefore the usefulness of the modal notion in question to increase the expressive capabilities of our language crucially depends on the presentation of the axiom system that shapes our syntactic/arithmetical world. I will elaborate more on this point in the next, final section.

6. A DILEMMA?

Let's pause for a moment and reconstruct in some more detail the main line of argumentation that may have been blurred by the necessary technicalities introduced in the previous sections. We started with arithmetical theories in their role of theories of the bearers of modal ascriptions. We then asked ourselves what are the sufficient conditions for characterizing the objects to which intensional notions apply: our answer led us to identify a class of arithmetical systems, the *sequential*

[14] See [5, 4.11] for a proof in the case of PA. The method can however be generalized to all B.

theories, as satisfactory candidates for such a role (§2). This is not an unconventional choice. Vann McGee, for instance, in his classic on truth writes that we require from a theory of the objects of truth

> '...the ability to describe expression types and their syntactic properties, the ability to talk about natural numbers and their mathematical properties, and the ability to talk about the coding relations.'([15, p. 18])

Via the interpretation of suitable syntactic theories, such as S^1_2, we can even require the provability of some universally quantified statements concerning the structure of syntactic objects.[15]

Once the properties of a suitable arithmetical/syntactical theory are identified, it is possible to formulate extensions our base theory with principles characterizing alethic modalities. We have considered principles that can be satisfied by a cluster of intensional notions, broadly conceived as to include truth and propositional attitudes: for a sequential base theory B, they amount to the *minimal theory* $\mathsf{M}[B]$, the theory of *grounded* modalities $\mathsf{M}[B]$, the theory of *compositional* modalities $\mathsf{C}[B]$. Admittedly, the principles characterizing the three theories are not aimed at pinpointing a single modal notion, but at being sufficiently general to capture the fine boundary between intensional notions that are *reducible* to the resources of the object theory, and intensional notions that are not.

What's the sense of *reducible* that is relevant here? We considered three possible choices: the *non conservativeness* (conservativeness) of the modal theory over the base theory, witnessing – following the tradition initiated by [23, 13] – the property of the modal theory of explaining (not being able to explain) non-modal facts concerning B not already explained by B; the relative interpretability (non-interpretability) of the modal theory in the base theory B, witnessing – as envisaged by [9, 5, 4] – the conceptual reducibility (irreducibility) of the modal notion to the inferential resources of B; the modal theory's having (not having) non-elementary speed up with respect to the base theory B, witnessing the expressive extra-power (or the lack of it) with respect to B ([5]).

The scenario that resulted from considering the reducibility of the three modal theories to a sequential theory B turned out to be at least puzzling. If conservativeness behaves uniformly for sequential theories extending a weak arithmetical system (i.e. EA), relative interpretability is highly dependent on the presentation

[15] This desideratum was clearly emphasized in Halbach's influential monograph:

> '... a base theory must contain at least a theory about the objects to which truth can be ascribed. [...] The axioms of the truth theory can serve their purpose only if the base theory allows one to express certain facts about syntactic operations; [...] it should be provable in the base theory that a conjunction is different from its conjuncts and different from any disjunction.[...] The decisive advantage of using Peano Arithmetic is that I do not have to develop a formal theory that can be used as base theory...' ([7, §2])

of the theory B, and speed up is likely to behave much more similarly to relative interpretability than to to speed up.

A natural thought would then be the following:

1. One should disregard notions of reduction that display a non-uniform conduct. In the case at hand, we should only take seriously the conservativeness (or non-conservativeness) of a modal theory over a sufficiently strong sequential theory. Given their idiosyncrasies, relative interpretability and speed up should not be trusted.

Let's reflect a bit on 1. Looking back at the philosophical theses associated with the three reducibility relations we have taken into account, a consequence of 1. would be to consider as legitimate the equation 'conservativeness = lack of explanatory power', whereas the conceptual reducibility of an intensional notion could not be coherently matched with the relative interpretability of the modal theory in the base theory, and similarly for the attempts of combining expressive power and non-elementary speed up – under the assumption of the truth of our conjectures above. The reason should be now clear but it's worth repeating in a direct and informal way: in constructing a modal theory we identified two main components. In the first place a satisfactory theory of the objects of modal ascriptions, that we identified with the class of sequential theories. Secondly, some principles governing some modal notion, that do not depend in any way from the choice of the underlying sequential theory we choose. However, we have seen that the very same principles characterizing a modal notion may be reducible or not reducible, expressively stronger or not, depending on which sequential theory we choose.

However, despite its initial plausibility, thesis 1. cannot be accepted. It is in fact clear that, in the study of the semantic or modal notions characterized by the axioms of $\mathsf{M}[B]$, $\mathsf{G}[B]$ and $\mathsf{C}[B]$, conservativeness is a coarser grained notion if compared to, for instance, relative interpretability. On the one hand, in fact, we have seen that there are cases in which $\mathsf{M}[B]$, $\mathsf{G}[B]$ and $\mathsf{C}[B]$ are conservative over B but not interpretable in B. On the other, by considering *natural and essential* extensions of, say, $\mathsf{G}[B]$ and $\mathsf{C}[B]$, we end up with theories that will all contain the results of extending the nonlogical schemata of $B \supseteq \mathsf{S}^1_2$ to the predicate $\Box$, we obtain the theories $\mathsf{G}[B]^+$ and $\mathsf{C}[B]^+$. Both theories will prove $\mathsf{Con}(B)$ and therefore will not be conservative over B by Gödel's second incompleteness theorem. However, this also entails, both in the case in which B is inductive and finite, that neither $\mathsf{G}[B]^+$ nor $\mathsf{C}[B]^+$ can be interpreted in B. In other words, for natural modal theories, relative interpretability implies conservativeness but not vice versa: *as a consequence, the study of the relative interpretability of the modal theory in the base theory adds substantial information about the notion in question that is not obtained via the question on the its conservativeness.* Although this is not a direct rejection of 1., I take it as a sign that a significant amount of important data would be lost once endorsing it.

A second reaction to the situation depicted in Tables 1-3 could be the following:

2. Although uniformity is lost if we consider *both* classes of finite and inductive theories, it is clearly regained if we focus on *each* of them. We must therefore choose between finite and inductive base theories. In particular, inductive theories feature an open-ended schema for the natural numbers that, given the close connection between the structure of syntactic and natural numbers, enables us to capture without arbitrary restrictions the intended domain of bearers of modal ascriptions. Therefore we must disregard finite theories.[16]

Also this option should be resisted, for reasons that are similar to the ones contained in my reaction to 1. If one is interested, for instance, in the conceptual reducibility of the notion captured by $\Box$, one would clearly realize that, for the very specific behaviour of relative interpretability in the context of inductive theories witnessed by Orey's theorem above and similar results, the relative interpretability of the modal theory in the inductive base theory cannot be a real test. It's only in the context of finite theories that these general collapsing phenomena between relations of interpretability are blocked, and the question of the conceptual reducibility of the modal predicate becomes relevant. It's there, and not elsewhere, that one can really see how the modal predicate helps in structuring the ontology of numbers (such as when we defined cuts on with the consistency of the base theory can be proved) in ways that are not otherwise available in the base theory. And it's also there that, as we have seen, the modal predicate *is* doing some conceptually irreducible work, and it's precisely this that would be hidden when we restrict our attention to inductive base theories only.

Similarly, under the assumption that the modal theories $\mathsf{G}[B]$ and $\mathsf{C}[B]$ have no significant speed up over an inductive B – a fact that would likely be extracted from the relative interpretability of $\mathsf{C}[B]$ in B for B inductive – one should conclude that the modal predicate in question has no expressive power or cannot be a useful tool to express facts concerning the syntactic base theory in a more concise and shorter way. However, by focusing on finite base theories, one suddenly realizes that there is a clear sense in which the modal notion in question *can* shorten proofs in the base theory: although it cannot cope with an infinite presentation of B, when the starting point is a finite set of axiom the modal predicate is indeed expressively stronger than any of the resources of B. In both cases, therefore, endorsing 2. and focusing only on inductive base theories would lead to a severe loss of significant information.

But are 1. and 2. two horns of a dilemma? In a sense, if one aims at clear-cut solutions that could result in a concise and effective philosophical message concerning all reductions at once, the answer is yes. The results mentioned above, however, also point at an irreducible level complexity that cannot be so easily resolved as it happens in both horns of the dilemma. One way to react to the deadlock is therefore to add an additional parameter to the evaluation of reduction of semantic and modal notions to an underlying base theory: the presentation of the base theory. In the

[16] A suggestion along these lines has been recently made by Fujimoto in [6].

light of what we said above, it is not sufficient to consider a modal theory as compounded by (i) a theory of bearers of modal ascriptions satisfying some minimal sufficient conditions and (ii) a set of principles characterizing a modal notion. We should also consider how the theory specified in (i) is presented to us: schematic, inductive, infinitely axiomatized, finitely axiomatized.

This conclusion, however, does not hold for *any* uses of modal theories: when only paradoxes and consistency are at stake, there is no need to consider subtle issues like the specific axiomatization of the truth bearers. But even outside the analysis of reductions of a modal theory to the base theory, for instance when one compares different principles belonging to a single solution to paradoxes, as Halbach and Nicolai do in [17] for Kripke's theory of truth, the presentation of the base theory is an non-eliminable parameter in measuring the proof-theoretic strength of such solutions.

References

[1] A. Cantini. Notes on formal theories of truth. *Zeitschrift für Logik un Grundlagen der Mathematik*, 35:97–130, 1989.

[2] S. Feferman. Arithmetization of metamathematics in a general setting. *Fundamenta Mathematicae*, 1960.

[3] Hartry Field. Deflating the conservativeness argument. *Journal of Philosophy*, 96(10):533–540, 1999.

[4] M. Fischer and L. Horsten. The expressive power of truth. *Review of Symbolic Logic*, 8(2):345–369, 2015.

[5] Martin Fischer. Truth and speed-up. *Review of Symbolic Logic*, 7(2):319–340, 2014.

[6] Kentaro Fujimoto. Deflationism beyond arithmetic. *Unpublished Manuscript*.

[7] V. Halbach. *Axiomatic theories of truth. Revised edition*. Cambridge University Press, 2014.

[8] Volker Halbach. Conservative theories of classical truth. *Studia Logica*, 62(3):353–370, 1999.

[9] L. Horsten. *The Tarskian Turn*. MIT University Press, Oxford, 2012.

[10] Paul Horwich. *Truth*. Clarendon Press, 1998.

[11] Petr Hájek and Pavel Pudlák. *Metamathematics of first-order arithmetic*. Perspectives in mathematical logic. Springer-Verlag, Berlin, New York, 1993.

[12] Thomas Jech. *Set Theory. The third millennium edition*. Springer, Berlin, 2008.

[13] Jeffrey Ketland. Deflationism and Tarski's paradise. *Mind*, 108(429):69–94, 1999.

[14] G. Leigh. Conservativity for theories of compositional truth via cut elimination. *The Journal of Symbolic Logic*, 80, 2015.

[15] V. McGee. *Truth, vagueness, and paradox*. MIT University Press, 1991.

[16] C. Nicolai. More on systems of truth and predicative comprehension. In F. Boccuni and A. Sereni, editors, *Objectivity, Realism, and Proof. FilMat Studies in the Philosophy of Mathematics*. Springer, 2016.

[17] C. Nicolai and V. Halbach. On the costs of nonclassical logic. *Journal of Philosophical Logic*, To appear.

[18] Carlo Nicolai. Deflationary truth and the ontology of expressions. *Synthese*, 192(12):4031–4055, 2015.

[19] Carlo Nicolai. A note on typed truth and consistency assertions. *Journal of Philosophical Logic*, 45(1):89–119, 2016.

[20] P. Pudlák. Some prime elements in the lattice of interpretability types. *Transactions of the American Mathematical Society*, 280:255–275, 1983.

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

[22] Pavel Pudlák. The lengths of proofs. In Samuel R. Buss, editor, *Handbook of Proof Theory*, volume 137 of *Studies in Logic and the Foundations of Mathematics*, pages 547 – 637. Elsevier, 1998.

[23] Stewart Shapiro. Proof and truth: Through thick and thin. *Journal of Philosophy*, 95(10):493–521, 1998.

[24] Stephen George Simpson. *Subsystems of second order arithmetic*. Perspectives in logic. Association for symbolic logic, New York, 2009.

[25] J. Stern. *Towards predicate approaches to modality*. Springer, 2016.

[26] A. Tarski. Der Wahrhetisbegriff in den formalisierten Sprachen. In *Logic, semantics, metamathematics*, pages 152–278. Clarendon Press, Oxford, 1956.

[27] A. Visser. What is the right notion of sequentiality. In C. Charampolas P. Cegielski and C. Dimitracopoulos, editors, *New Studies in Weak Arithmetics, volume 211 of CSLI Lecture Notes*, pages 229–272. Publications and Presses Universitaires du Pole de Recherche et d'Enseingement Superieur Paris-est, Stanford, 2013.