PHILOSOPHICAL PERSPECTIVES

Philosophical Perspectives, 32, Philosophy of Language, 2018
doi: 10.1111/phpe.12113

DEFLATIONISM AND THE FUNCTION OF TRUTH

Lavinia Picollo
University College London

Thomas Schindler
University of Amsterdam

1. Introduction

Deflationism about truth is a trending cluster of claims about the meaning and the role of the truth predicate in natural language, as well as the nature of truth itself. Different versions of deflationism pick different claims as part of the core and belt of the position. The most popular version, which will be the focus of this paper, takes the two theses we describe next to constitute its core. Its most vocal contemporary advocates are Horwich and Field.

The first core thesis is that the content of a truth ascription to a given sentence or proposition can be entirely articulated in terms of the equivalence between the ascription and the sentence or proposition itself. Thus to assert that 'snow is white' is true, or that the proposition that snow is white is true, is simply to assert that show is white. We will often refer to this as the 'equivalence thesis'. It has its origins in the redundancy theory of truth, a proto-deflationism advanced by Frege, Ramsey, and Ayer.

The second core thesis concerns the role that the truth predicate plays in natural language. Like the redundantists, deflationists hold that, despite appearing grammatically as an adjective, the role of the truth predicate is not descriptive. But unlike redundantists, deflationists argue that it has an important expressive function. For example, it allows for blind ascriptions such as 'Goldbach's conjecture is true' and generalisations that otherwise would be very hard or perhaps impossible to express, such as 'All theorems of arithmetic are true'. According to deflationism, this function is the sole reason for the existence of a truth predicate in natural language.

Oddly enough, in spite of its central role in deflationism, the expressive function of the truth predicate has rarely been subjected to a rigorous investigation. This is highly regrettable because there are many controversial issues

concerning the interpretation of the deflationary account of truth that are, unsurprisingly, directly or indirectly affected by one's view on this function. An important question is whether, in light of the semantic paradoxes that emerge within classical truth theories, deflationists are compelled to adopt non-classical logics; and whether these logics allow the truth predicate to perform its expressive function. Another issue, which we address in section 5, revolves around the so-called conservativity requirement. Some deflationists have claimed that truth is an insubstantial, metaphysically thin, or lightweight notion. This is often taken to imply that truth cannot yield new knowledge about the world or have any explanatory power. What this precisely amounts to has remained largely unclear. A number of philosophers have suggested that this impossibility binds deflationism to the conservativity of their truth theories. It has been argued that adding a deflationary truth predicate to a theory must yield a conservative extension of the latter, which means that every sentence not containing the truth predicate that is derivable in the extended theory must already be derivable in the original theory. This generates a conflict, because most interesting theories of truth do not lead to a conservative extension.

In this paper we put forward a precise analysis of the function of truth. This analysis has interesting implications on several matters concerning deflationism, including the conservativity issue and the question whether deflationists should reject classical logic, which we leave for later work. The central claim of this paper is that the truth predicate is best understood as a tool for *mimicking higher-order quantification* in a first-order framework. In other words, the truth predicate allows us, in an indirect way, to quantify into sentence and predicate position without introducing sentential or predicate quantifiers.

Actually, this is not an entirely novel claim, and several results concerning the intertranslatability between languages with higher-order quantifiers and languages with a truth predicate have emerged in the more formal literature on truth. However, these results are partial and limited in scope, and therefore have not lead to a clear understanding of the function of truth. First, they often relate *predicative* fragments of higher-order logic to theories of truth, leaving the wrong impression that impredicative higher-order quantification cannot be handled by a truth predicate. Second, they often relate these fragments of higher-order logic to *compositional* theories of truth, suggesting that they cannot be handled by a disquotational truth predicate. Third, in the philosophical literature there is an almost exclusive focus on the relation between *sentential* quantifiers and the truth predicate, neglecting the fact that the truth predicate actually allows us to mimic quantification into predicate position as well.

We provide formal results showing that (i) languages with impredicative sentential quantifiers can be translated into languages containing a disquotational truth predicate and vice versa, and that (ii) languages with impredicative predicate quantifiers can be translated into languages containing a disquotational truth predicate as well and vice versa. In particular, one can show that the whole simple theory of types is reducible to disquotational principles of truth.

Moreover, our results lead to natural answers concerning the conservativity debate. Roughly, if we are to understand the truth predicate as a tool for simulating higher-order quantification in a first-order framework, we should not expect that the addition of a truth predicate leads to a conservative extension because, in general, the addition of higher-order quantifiers does not lead to a conservative extension either.

This paper is organised as follows. In section 2 we briefly review and criticise some accounts of the expressive function of truth that one can find in the literature, and motivate our own approach, which we back up with formal results in sections 3 and 4. In section 3 we will show that languages with sentential quantifiers can be translated into languages containing a truth predicate and vice versa. In section 4 we will show that languages with predicate quantifiers can be translated into languages containing a truth predicate and vice versa. Since this paper is intended for a general philosophical audience, technicalities will be kept to a minimum. Formal details and proofs of our results can be found in our forthcoming companion paper "Truth and Higher-Order Quantification". In the penultimate section of the paper, we turn to the conservativity debate, arguing, as indicated above, that deflationists are not committed to the conservativity of their theories of truth. In the final section, we point to a number of further philosophical consequences of our results. A proper discussion of these consequences must be relegated to a separate paper.

2. The Function of the Truth Predicate

Grammatically, the word 'true' works as an adjective in natural language. In the cases we are interested in it applies to sentences, propositions, statements, or other similar kinds of objects with propositional content. Accordingly, in the logical structure of sentences containing truth ascriptions, 'true' is in the vast majority of cases taken to be a predicate, with a few notable exceptions (cf. Grover *et al.* [9]). In this paper we follow the tradition.

Being a predicate, 'true' applies to noun phrases. When it's attached to a transparent name, i.e. to a name that allows us to read off the truth bearers they are names of, as in "'Caesar was murdered" is true' or 'It is true that Caesar was murdered', we say that the truth ascription is *explicit*. In those cases, the truth predicate seems redundant, for the same information could have been conveyed without it, simply by uttering 'Caesar was murdered'. It looks like the word 'true' is just being used for *emphasis*; it has merely a rhetorical or stylistic function.

By contrast, when the truth predicate is attached to a non-transparent name, its function can go beyond a matter of style. Although sentences like

Goldbach's conjecture is true (1)

are equivalent to the sentence they ascribe truth to, the truth predicate might not be as redundant and easily eliminable as before. For one might not be in a position to *explicitly articulate* Goldbach's conjecture, e.g. because one doesn't know its content, or because it would take too long to do so. Thus, it is often said that the truth predicate has an epistemic use, as well as a lazy use. Truth ascriptions like (1) are usually referred to as 'implicit' or 'blind' ascriptions. Similar effects can be achieved by a different kind of implicit ascription, in which the truth predicate is attached to general terms instead. They are often called 'generalisations' or 'general truth ascriptions'. These are, for instance,

The last thing Tarski said about truth is true

where a definite description is used to refer to the object we wish to ascribe truth to, or

Everything Tarski said about truth is true (2)

where a predicate is employed to restrict the class of objects we say are true. The last example differs from the previous ones in that truth is ascribed not just to one but to many sentences or propositions at once.

All the examples we have considered so far involve *finite truth ascriptions*, that is, truth is ascribed only to finitely many objects. When blind, these ascriptions are frequently described as devices for *indirect endorsement* or *indirect assertion*. However, none of these descriptions adequately reflects the nature of the function that the truth predicate is playing here. Endorsing or asserting the content of a sentence is just one among many other things one can do with a truth ascription of this kind. Certainly, one can endorse Goldbach's conjecture without explicitly articulating it, by affirming (1). But one can also *deny* the conjecture, by negating (1), *assume* it, by assuming (1), *state a necessary condition* for it, by embedding (1) in the antecedent of a conditional, etc., all without explicitly articulating Goldbach's conjecture. For similar reasons it would be misleading to limit the function of truth in blind ascriptions to the expression of agreement and disagreement. In blind ascriptions the truth predicate serves the more general purpose of allowing us to blindly or indirectly *express* the content of one or many sentences, without explicitly articulating these sentences.

This is also true of generalisations in which truth is ascribed to *infinitely* many objects, as in

All theorems of arithmetic are true (3)

or

Some theorems of arithmetic are true (4)

We may call these generalisations 'infinite truth ascriptions'. Although they can also serve an epistemic function – for instance, when we don't know the infinitely many sentences we wish to ascribe truth to – they can hardly be said to have a lazy use. It's not out of laziness of explicitly articulating, e.g. the infinitely many theorems of arithmetic that we may use a sentence such as (3), but due to the actual impossibility of explicitly articulating them all. This impossibility entails an *indispensability* of some sort, that singles out the role of truth in infinite truth ascriptions as the most important one from a logical standpoint. In Quine's famous words,

> We may affirm the single sentence by just uttering it, unaided by quotation or by the truth predicate; but if we want to affirm some infinite lot of sentences that we can demarcate only by talking about the sentences, then the truth predicate has its use. (Quine [27, p. 12])

In what sense does the truth predicate allow us to affirm (or express) an infinite lot sentences? A popular account has it that infinite truth ascriptions are equivalent to infinite conjunctions and infinite disjunctions (e.g. Putnam [26, p. 15)], Gupta [10, pp. 60-61], Horwich [20, p. 3)]). For example, Putnam writes:

> If we wanted to say 'what he said is true', for example, we could say:
>
> (1) [He said 'p_1' and p_1] or [He said 'p_2' and p_2] or ...
>
> where the disjunction in (1) contains one disjunct for each sentence 'p_i' of the object language. But we *can't*, as a matter of fact, speak in infinite disjunctions. So instead we look for a finite expression equivalent to (1). Now,
>
> (2) For some x he said x & x is true
>
> will be equivalent to (1) provided for each i ($i = 1, 2, 3, \ldots$)
>
> (3) 'p_i' is true if and only if p_i
>
> is correct. But this is just Tarski's 'Criterion T' [...] [26, p. 15]

Similarly, this account has it that a generalisation like (3) is equivalent to the infinite conjunction '[If 'p_1' is a theorem of arithmetic then p_1] and [If 'p_2' is a theorem of arithmetic then p_2] and ...'

There is a sense in which this is correct: under the *intended interpretation* of the language, generalisations like (3) and the set of its instances have the same truth conditions, because the range of the quantifier is precisely the class of sentences of the language under consideration. From a logical point of view, however, understanding these generalisations as infinite conjunctions is as misguided as understanding any general claim as the infinite conjunction of its instances. Quantifiers and infinite conjunctions serve the common purpose of

collecting infinite information into one single expression, but do so in different ways. Their inferential behaviour diverges in several respects, and so does their meaning. For one thing, quantified claims may belong themselves to the range of their own quantifiers, whereas infinite conjunctions and disjunctions are not allowed to have themselves as proper subformulae. For another, unlike infinite conjunctions, universal generalisations are not generally entailed by the class of their instances, due to compactness considerations.

In light of these issues, Halbach [11], [14] has suggested an alternative way of understanding Quine's talk of 'affirming some infinite lot of sentences'. In order to avoid problems with the semantic paradoxes, Halbach limits his attention to sentences that do not contain the truth predicate themselves, that is, he works with a restricted version of the T-schema. Consider sentence (3) again. Given the (restricted) T-schema, (3) implies all sentences of the form:

If 'p' is a theorem of arithmetic then p (5)

Of course, (3) is not the only sentence that will imply them; any contradiction will do the same job. But, as Halbach observes, (3) implies no sentence in the language without the truth predicate that does not already follow from some instances of (5). In other words, given the T-schema, (3) and the set of all instances of (5) are *equivalent in their truth-free consequences*. One can say, then, that the general truth ascription (3) *finitely axiomatises* the infinitely many instances of (5) relative to the T-schema.

While this is an insightful observation, it falls short of providing a full analysis of the role the truth predicate plays in infinite truth ascriptions. First, as we observed in [25], the equivalence of the truth-free consequences of generalisations like (3) and those of the set of instances of (5) can be established using only one direction of the T-schema, namely, its elimination half,

If 'p' is true, then p (T-Out)

That is, (3) finitely axiomatises the set of sentences of the form (5) relative to T-Out alone. This seems to indicate that there is something missing in Halbach's account. Second, the account is silent on the expressive function of truth when generalisations like (3) are embedded in a conditional or conjoined with another sentence.[1] Third, an account of the role of the truth predicate in infinite generalisations should also be able to explain its utility if we want to generalise on sentences that contain the truth predicate themselves. A natural extension of Halbach's approach to cases in which the sentences p may contain the truth predicate themselves would require that the general truth ascription and the set of its 'instances' have the same consequences *simpliciter*. Since the general truth ascription is a consequence of itself, this means the set of its 'instances' would need to imply the generalisation as well. However, again, this is not generally to be expected.

But then what role is the truth predicate playing in infinite truth ascriptions? The following quote by Quine is insightful:

> We can generalise on 'Tom is mortal', 'Dick is mortal', and so on, without talking of truth or of sentences; we can say 'All men are mortal'. We can generalize similarly on 'Tom is Tom', 'Dick is Dick', '0 is 0', and so on, saying 'Everything is itself'. When on the other hand we want to generalize on 'Tom is mortal or Tom is not mortal', 'Snow is white or snow is not white', and so on, we ascend to talk of truth and of sentences, saying 'Every sentence of the form 'p or not p' is true', or 'Every alteration of a sentence with its negation is true'. What prompts this semantic ascent is not that 'Tom is mortal or Tom is not mortal' is somehow about sentences while 'Tom is mortal' and 'Tom is Tom' are about Tom. All three are about Tom. We ascend only because of the oblique way in which the instances over which we are generalizing are related to one another. [27, p. 11]

We can learn two lessons from this passage. First, the truth ascription 'All sentences of the form "p or not p" are true' is to the instances of the law of excluded middle what 'All men are mortal' is to its corresponding instances, that is, 'Tom is mortal', 'Dick is mortal', and so on. Thus, the relation between an infinite truth ascription and its corresponding 'instances' is akin to the relation, not between a conjunction and its conjuncts, or between an axiomatisation and the sentences it axiomatises, but between a universal claim and its instances.

Second, and more importantly, the truth predicate allows us (indirectly) to generalise on a syntactic position that first-order quantifiers are incapable of, namely, that of a sentence. We cannot replace the clause 'Tom is mortal' in 'Tom is mortal or Tom is not mortal' with a pronoun (or variable) – or rather, we cannot do so in English (or in a first-order language). Of course, this is just a contingent fact about English. We can imagine an alternative history in which English developed in such a way that pronouns can occur in sentential positions while still referring back to the quantifiers, and we can devise artificial languages that do contain sentential variables and quantifiers. The point is that such quantifiers can be dispensed with in the presence of a truth predicate. Given that each sentence p is equivalent to "'p" is true', we can substitute 'Tom is mortal' with "'Tom is mortal" is true', yielding

> 'Tom is mortal' is true or 'Tom is mortal' is not true

Since in the displayed sentence the clause 'Tom is mortal' is mentioned rather than used, we can replace the term by a variable, yielding 'x is true or x is not true', and that in turn allows us to formulate the generalisation 'For every sentence x, x is true or x is not true'. Thus, the truth predicate allows us, in combination with our ordinary quantifiers, to quantify (indirectly) into sentence position – circumventing the need to introduce sentential quantifiers, i.e. quantifiers that directly bind sentence places in our language. *This is the nature of the role that the truth predicate plays both in finite and infinite truth ascriptions*

according to deflationism. The truth predicate is a useful syntactic restructuring device. Its value stems from the ungrammatical character of sentential quantification in English, and the ubiquity of first-order languages in science and philosophy. We may say, following Horwich [20, p. 125], that 'the value of our concept of truth [is] that it provides, *not the only way*, but a relatively 'cheap' way of obtaining the problematic generalizations– the way actually chosen in natural language' (our emphasis).

The idea that the truth predicate allows for quantification into sentence position was anticipated by Ramsey, who claimed that truth could be eliminated if sentential quantifiers were in our toolbox [28, p. 158]. Akin ideas can also be found, for instance, in Tarski [33] (cf. Carnap [4, p. 50]), where a definition of truth in terms of sentential quantifiers is considered (and ruled out for technical reasons); in Grover [8] and Grover *et al.* [9], who equate sentential quantifiers and the truth predicate by ascribing both a prosentential role; and in Azzouni [2], [3], who replaces the truth predicate with a special kind of quantifier which simultaneously binds variables in sentential and nominal position.

In the next section we will present a formal result that languages with sentential quantifiers can be translated into languages with a disquotational truth predicate and vice versa. This translation is novel as it preserves not only theorems but also *inferences*, as well as it improves upon earlier results in that it shows that the truth predicate can even handle *impredicative* sentential quantification, i.e. quantification of sentences that contain themselves sentential quantifiers.

However, mimicking sentential quantification does not exhaust the capacity of a disquotational truth predicate: it also allows us to mimic quantification into predicate position. This often overlooked fact has far-reaching consequences. Consider again the sentence

Tom is mortal (6)

First-order languages do not allow us to generalise on the syntactic position of predicates, and natural languages such as English do not allow us to do so in general either. Again, this is just a contingent fact about English. If we want to generalise predicate places, we can turn to some artificial framework such as the language of second-order logic or its infinite iteration, the simple theory of types. In such languages, we can formulate sentences such as $\exists X\, X(\text{Tom})$. But we can turn to the truth predicate as well to achieve a similar effect. Disquotation implies that (6) is equivalent to

'Tom is mortal' is true

Since the sentence 'Tom is mortal' is the result of concatenating the name 'Tom' with the predicate 'is mortal', (6) is also equivalent to

The concatenation of 'Tom' with 'is mortal' is true

In the last formula, the predicate 'is mortal' is mentioned rather than used, so we can replace the term with a variable – allowing us to say, for example,

There is a predicate x such that the concatenation of 'Tom' with x is true

This statement corresponds in an obvious way to the second-order sentence $\exists X\, X(\text{Tom})$.

In the section 4 we will present a formal result that languages with *impredicative* predicate quantifiers can be translated into languages with a disquotational truth predicate and vice versa. But before moving on to the next section, there is a restriction that is worth pointing out. Truth theories can handle quantification into predicate position by talking about the concatenation of certain terms with certain formulas being true, as indicated above. But this means that truth theories can only simulate predicate quantification for those languages that contain a name for each individual in the domain of the first-order quantifiers – thus, the domain needs to be countable. However, this restriction can be dropped if we operate with a satisfaction predicate instead of a truth predicate.

3. Truth and Sentential Quantification

We proceed with the presentation of our formal results connecting truth predication and quantification into sentence position. As this paper is intended for a general philosophical audience, we will keep our exposition as simple as possible, deferring a detailed presentation of our results to our companion paper "Truth and Higher-Order Quantification". However, technicalities cannot be avoided altogether: technically less inclined readers may skim through the following material, focusing merely on the definition of the higher-order theory QPL, the definition of the truth theory UTB[τ], and Theorem 1, which relates the two.

Let $\mathcal{L}$ be a first-order language with the usual logical vocabulary including identity and possibly enumerably many non-logical operators $\Box_1, \ldots, \Box_n, \ldots$ We will be interested in several extensions of $\mathcal{L}$. The first extension, $\mathcal{L}_{\mathrm{Q}}$, is the result of augmenting $\mathcal{L}$ with sentential variables α_i for each natural number i. Every sentential variable α_i is a formula of $\mathcal{L}_{\mathrm{Q}}$ and, if φ is a formula of $\mathcal{L}_{\mathrm{Q}}$, so is $\forall\alpha_i\, \varphi$. The existential sentential quantifier can be defined in terms of the universal quantifier and negation, as usual. Sometimes we write $\alpha, \beta, \ldots$ instead of $\alpha_0, \alpha_1, \ldots$, for readability.

We extend classical first-order logic with sentential quantifiers by adding the following rules to the calculus:

$$(\forall\alpha\mathrm{I})\ \frac{\varphi}{\forall\alpha_i\, \varphi} \qquad\qquad (\forall\alpha\mathrm{E})\ \frac{\forall\alpha_i\, \varphi}{\varphi[\psi/\alpha_i]}$$

provided that α_i in ($\forall\alpha$I) is arbitrary and ψ is free for α_i in φ in ($\forall\alpha$E). In ($\forall\alpha$E), ψ can be any formula of the language without free individual variables, as sentential variables are intended to range over sentences. ψ may, however, contain any number of bound individual variables and any number of free and bound sentential variables. Let QPL be the extension of first-order logic with the rules ($\forall\alpha$I) and ($\forall\alpha$E), plus Necessitation and the K-axiom for each of the operators $\Box_1, \ldots, \Box_n$ $\Box_1, \ldots, \Box_n$.

It is easy to see that ($\forall\alpha$E) entails the following comprehension principle:

$$\exists\alpha\,(\alpha \leftrightarrow \varphi) \qquad \text{(Sentential comprehension schema)}$$

where α doesn't occur free in φ. Again, φ may contain any number of bound individual variables and any number of free and bound sentential variables, though no free individual variables. In other words, the Sentential comprehension schema is highly impredicative, and so is ($\forall\alpha$E). Of course, if one preferred to work within a predicative framework, one could simply restrict ($\forall\alpha$E) to formulas ψ containing no bound sentential variables (and perhaps no free sentential variables either). In this system the Sentential comprehension schema will also be restricted to such expressions.

We can turn to QPL to generalise into sentence position in the expected way. For instance, we can express that every sentence of the form $\varphi \vee \neg\varphi$ is true, as in Quine's example, by the formula $\forall\alpha\,(\alpha \vee \neg\alpha)$. Here, α can be instantiated with any formula of $\mathcal{L}_{\mathrm{Q}}$ (not containing free individual variables), including the sentence $\forall\alpha\,(\alpha \vee \neg\alpha)$ itself. We might also want to express more contingent matters, e.g. that everything Angela Merkel said about the refugee crisis is true. To do so we may take a monadic operator $\Box$ of language to apply only to sentences about the refugee crisis uttered by Angela Merkel (and their logical consequences) and write $\forall\alpha\,(\Box\alpha \rightarrow \alpha)$.[2] Note that the same effect can hardly be achieved using predicates instead of operators, for sentential variables are formulas and predicates can only be syntactically attached to terms.[3]

Since we want to relate languages with sentential quantifiers to languages with a truth predicate, let us proceed by describing the latter. As truth applies to specific objects, the truth bearers, formal principles of truth are normally not formulated directly over logic, but over an underlying theory that has names for those objects and proves facts about them. As is usual in the literature on formal theories of truth, we choose sentences as truth bearers. Should propositions be preferable to sentences, one could understand our truth predicate as applying not directly to the sentences but to what these sentences express. As a consequence, our underlying system will be a theory of syntax, which we assume to be sufficiently rich to be able to interpret a decent amount of arithmetic, as is usual – interestingly, we need not assume that principles of induction are available in the syntax theory. Let $\mathcal{L}_{\mathrm{T}}$ be the result of augmenting $\mathcal{L}$ with unary predicates T, for truth, and D, for the domain of objects $\mathcal{L}$ talks about, and the vocabulary needed to formulate a theory of syntax, Σ, for the whole language, including a

unary predicate S for the domain of the syntax theory. To keep the ontology of $\mathcal{L}$ apart from the syntactic objects Σ brings about, we assume all axioms in Σ are relativised to S.

It is possible to show that there exists an effective translation, τ, of expressions of $\mathcal{L}_{\mathrm{Q}}$ into expressions of $\mathcal{L}_{\mathrm{T}}$ that has the following natural properties. First, atomic formulas of the form $Pt_1 \ldots t_n$ are translated into $\bigwedge \mathrm{D}(\vec{t}) \rightarrow Pt_1 \ldots t_n$ and their negations into $\bigwedge \mathrm{D}(\vec{t}) \rightarrow \neg Pt_1 \ldots t_n$, where D (as mentioned above) is a predicate applying to all and only those objects in the domain of things that $\mathcal{L}$ talks about, and $\bigwedge \mathrm{D}(\vec{t})$ is short for $\mathrm{D}(t_1) \wedge \ldots \wedge \mathrm{D}(t_n)$ (and similarly for other predicates). This relativisation is required because the quantifiers of the target language $\mathcal{L}_{\mathrm{T}}$ range over both the objects of $\mathcal{L}$ and those of the syntax theory.

Second, sentential variables α are translated into $\mathrm{Trsl}(x) \rightarrow \mathrm{T}x$ and their negations into $\mathrm{Trsl}(x) \rightarrow \neg \mathrm{T}x$. Here, $\mathrm{Trsl}(x)$ is a predicate of $\mathcal{L}_{\mathrm{T}}$ that applies exactly to (the codes of) those formulas of $\mathcal{L}_{\mathrm{T}}$ that are translations of formulas of $\mathcal{L}_{\mathrm{Q}}$ under τ.[4] Again, this relativisation is required because as $\mathcal{L}_{\mathrm{T}}$ contains the vocabulary of the syntax theory, its quantifiers range over more sentences than the sentential quantifiers of the original language $\mathcal{L}_{\mathrm{Q}}$. In order to preserve all inferences, the truth predicate only needs to apply to (the translation of) sentences of the original language. However, putting the relativisation aside, note that our translation replaces each sentential variable with a truth ascription to a first-order variable, restructuring expressions just as we anticipated in the previous section. Soon we will show how this truth predicate, combined with ordinary quantifiers and disquotational principles, actually allows us to quantify into sentence position just like sentential quantifiers.

Finally, the translation of complex formulas is determined by their logical structure: for example, $\varphi \wedge \psi$ is translated as $\tau(\varphi) \wedge \tau(\psi)$, while $\neg(\varphi \wedge \psi)$ is translated as $\tau(\neg\varphi) \vee \tau(\neg\psi)$. $\forall\alpha\, \varphi$ is translated as $\forall x\, \tau(\varphi)$, while $\neg \forall\alpha\, \varphi$ is translated as $\exists x\, \tau(\neg\, \varphi)$.[5]

We now introduce a truth theory that is able to simulate the logic of QPL. Let $\mathsf{UTB}[\tau]$ be the truth theory formulated in $\mathcal{L}_{\mathrm{T}}$ that results from extending the syntax theory Σ with all instances of the following schema:

$$\forall \vec{x} \left(\bigwedge \mathrm{Trsl}(\vec{x}) \rightarrow (\mathrm{T}^{\ulcorner}\varphi(\dot{\vec{x}})^{\urcorner} \leftrightarrow \varphi(\vec{x})) \right) \qquad (\mathrm{UTB}[\tau])$$

where φ is a formula in the range of τ. Here, $\vec{x}$ is short for $x_1, \ldots, x_n$ and the dot in $\mathrm{T}^{\ulcorner}\varphi(\dot{\vec{x}})^{\urcorner}$ indicates that the variables $\vec{x}$ are bound by the quantifiers $\forall\vec{x}$.

The reason why the variables $\vec{x}$ are relativised to the predicate Trsl is straightforward: the formula φ is a translation of some formula ψ of $\mathcal{L}_{\mathrm{Q}}$. The variables $\vec{x}$ correspond to the free sentential variables $\vec{\alpha}$ that may occur in ψ. Since the sentential variables $\vec{\alpha}$ range over sentences of $\mathcal{L}_{\mathrm{Q}}$, the corresponding variables $\vec{x}$ need to range over the translation of sentences of $\mathcal{L}_{\mathrm{Q}}$, if our translation is

to preserve inferences. That this is indeed so is the content of the following fundamental result of this section.

> **Theorem 1** If $\varphi_1, \ldots, \varphi_n, \psi$ are formulas of $\mathcal{L}_{\mathrm{Q}}$ then there is a derivation of ψ from $\varphi_1, \ldots, \varphi_n$ in QPL if and only if $\tau(\psi)$ is derivable from $\tau(\varphi_1), \ldots, \tau(\varphi_n)$ in UTB[τ].

This shows that any inference that we can draw using sentential quantifiers can be mimicked in a language with a disquotational truth predicate in a fairly natural way, that is, roughly as it was described in the previous section. This means one can use our theory of truth as a higher-order logic. In particular, if Γ is a theory formulated in $\mathcal{L}_{\mathrm{Q}}$, then ψ is a theorem of Γ if and only if $\tau(\psi)$ can be derived from $\tau(\Gamma)$, i.e. the set of translations of all non-logical axioms of Γ. Theorem 1 also entails the consistency of our truth theory UTB[τ]. Moreover, one can show that UTB[τ] is sound, in the sense of having a standard model for the underlying syntax theory (see our companion paper [Authors] for a proof).

Note also that an analogous result holds between predicative fragments of QPL, where the Sentential comprehension schema is restricted to formulae not containing sentential quantifiers (and possibly no free variables), and subsystems of UTB[τ] in which the instances of disquotation are restricted to the corresponding translations. In the specific case where no free variables are allowed in the instances of the comprehension schema, the corresponding theory of truth will be a typed system. However, as soon as free variables are permitted, the truth system needs to be type-free.

We do not necessarily recommend UTB[τ] as the ultimate truth theory, as it's rather artificial, being parasitic on $\mathcal{L}_{\mathrm{Q}}$. However, it is good enough to make our philosophical point. Having instances of disquotation for every translation allows us to *mimic* quantification into sentence position ranging over sentences of $\mathcal{L}_{\mathrm{Q}}$. This point can be generalised in an obvious way. If a truth theory contains instances of (uniform) disquotation for a certain class of sentences, one may say it enables quantification into sentence position ranging over this class. This, we would like to suggest, is the nature of the function of truth deflationists have in mind when they claim the truth predicate allows us to 'affirm some infinite lot of sentences'.

Of course, one has to be careful not to fall prey to paradoxes. To quantify over absolutely every sentence of the language including those containing the truth predicate, an unrestricted version of disquotation is required. Since in classical logic this principle leads to contradiction in combination already with a fairly weak syntax theory, one must adopt a weaker logic instead. However, note that not every logical system that guarantees consistency would do. One must take care that the rules and principles required to allow the truth predicate to perform its function still hold and, further more, one should ponder costs and benefits of adopting a weaker notion of logical consequence.

4. Truth-of and Higher-order Quantification

As indicated at the end of section 2, the function of the truth predicate is not exhausted by its ability to simulate sentential quantification. It also allows us to simulate quantification into predicate position as well. However, we mentioned a limitation: the truth predicate allows us to handle predicate quantification for any first-order theory whose intended interpretation has a countable domain but not for theories with an uncountable domain. This restriction could be dropped if we worked with a satisfaction predicate instead.

As before, let $\mathcal{L}$ be a first-order language and let $\mathcal{L}_2$ extend $\mathcal{L}$ with n-place predicate variables X_i^n for each natural number n and i. If $t_1, \ldots, t_n$ are individual terms, $X_i^n t_1 \ldots t_n$ is a formula of $\mathcal{L}_2$, and if φ is a formula of $\mathcal{L}_2$ then so is $\forall X_i^n \varphi$. We often drop the indices and write $X, Y, \ldots$ instead of X_i^n, for perspicuity. Second-order logic (SOL) extends first-order logic with the following axioms:

(A5) $\forall X^n \varphi \rightarrow \varphi[T/X^n]$, if T is an n-ary relation variable free for X^n in φ or an n-ary

(Comprehension schema) $\exists X^n \forall \vec{x}(X^n \vec{x} \leftrightarrow \varphi(\vec{x}))$, if X^n is not free in φ

(Second-order generalisation) $\frac{\varphi \rightarrow \psi}{\varphi \rightarrow \forall X \psi}$, if X is not free in φ or in a premise.

In the Comprehension schema, as in its sentential version, φ may contain free variables (other than $x_1, \ldots, x_n$). But in this case the variables can be both first- and second-order, for second-order quantifiers allow us to quantify into predicate position.

As before, we can consider (predicative) fragments of SOL in which the instances of the comprehension schema are restricted to formulas φ of a particular form (e.g. formulas containing no bound predicate variables).

Warning. When using the phrase 'second-order logic', some authors refer to the second-order logical consequence relation that is given by the so-called standard semantics for second-order languages, where the predicate variables are interpreted to range over the *full* power set of the domain. However, unless otherwise stated, in this paper 'second-order logic' simply refers to the calculus described above, which is sound and complete with respect to the class of so-called faithful Henkin models.[6] Thus, we understand second-order logic here as a theory of quantification into predicate position, and not as a theory of pluralities, properties, classes, sets, or concepts.

$\mathcal{L}_2$ allows us to bind predicate places in the language directly, as we mentioned in section 2. For instance, one can turn to second-order quantifiers to provide a finite axiomatisation of the principle of mathematical induction, as follows:

$$\text{(Induction axiom)} \quad \forall X((X0 \wedge \forall x(Xx \rightarrow X(x+1))) \rightarrow \forall x\, Xx)$$

The second-order theory that consists of the axioms of first-order Peano Arithmetic (PA) except that the usual Induction schema is replaced with the Induction

axiom is known as 'second-order arithmetic' or Z_2. In what follows we show that one can translate second-order variables and quantifiers employing a truth predicate pretty much as indicated at the end of section 2 in such a way that all inferences are preserved if suitable disquotational principles are available. In particular, one could formulate a system of arithmetic as strong as Z_2 in a theory of truth instead.

Indeed, using similar techniques as in the previous section, it is possible to show that there is an effective translation, σ, from expressions of $\mathcal{L}_2$ to expressions of $\mathcal{L}_\mathrm{T}$ that satisfies some natural properties. First, atomic formulas for the form $Pt_1 \ldots t_n$ (where P is a predicate symbol) and their negations are translated just as before, relativising the terms to $\mathrm{D}(x)$– a predicate applying to all and only those objects in the domain of things that $\mathcal{L}$ talks about. Second, we can again find a predicate $\mathrm{Trsl}(x)$ of $\mathcal{L}_\mathrm{T}$ that applies exactly to those formulas of $\mathcal{L}_\mathrm{T}$ that are translations of formulas of $\mathcal{L}_2$ under σ. Formulas of the form Xt are then translated into

$$\mathrm{D}(t) \wedge \mathrm{Trsl}(x) \rightarrow \mathrm{Ts}(x, t)$$

where s represents the function that, applied to a formula $\varphi(v)$ and a term t, yields the formula $\varphi[t/v]$. In more informal terms, σ translates Xt as 'If t is in the domain of $\mathcal{L}$ and x is the translation of a formula of $\mathcal{L}_2$, then x is true of t'. Similarly, formulas of the form $\neg Xt$ are translated as

$$\mathrm{D}(t) \wedge \mathrm{Trsl}(x) \rightarrow \neg\mathrm{Ts}(x, t)$$

Formulas of the form $Xt_1 \ldots t_n$ and $\neg Xt_1 \ldots t_n$ are dealt with in an analogous way. As before, the translation of complex formulas is determined by the logical structure of a formula.

Let $\mathsf{UTB}[\sigma]$ be the truth theory formulated in $\mathcal{L}_\mathrm{T}$ that results from extending the syntax theory Σ with all instances of the following schema:

$$\forall\vec{x}\forall\vec{y}\,(\bigwedge \mathrm{D}(\vec{y}) \wedge \bigwedge \mathrm{Trsl}(\vec{x}) \rightarrow (\mathrm{T}^{\ulcorner}\varphi(\dot{\vec{x}}, \dot{\vec{y}})^{\urcorner} \leftrightarrow \varphi(\vec{x}, \vec{y}))) \qquad (\mathrm{UTB}[\sigma])$$

where φ is a formula in the range of σ.

Again, the reason for the antecedent in the axioms is straightforward: φ must be the translation of some formula ψ of $\mathcal{L}_2$. ψ may contain free individual and free predicate variables. The variables $\vec{y}$ correspond to the free individual variables and the variables $\vec{x}$ correspond to the free predicate variables. Since the predicate variables of in ψ range over predicates of $\mathcal{L}_2$, the variables $\vec{x}$ need to range over the translation of these predicates. Similarly, since the free individual variables of ψ range over the domain of things that $\mathcal{L}$ talks about, the variables $\vec{y}$ need to be relativised to the predicate D, which applies exactly to those objects.

We can now state the main result of this section:

Theorem 2 If $\varphi_1, \ldots, \varphi_n, \psi$ are formulas of $\mathcal{L}_2$ then there is a derivation of ψ from $\varphi_1, \ldots, \varphi_n$ in SOL if and only if $\sigma(\psi)$ is derivable from $\sigma(\varphi_1), \ldots, \sigma(\varphi_n)$ in $\mathsf{UTB}[\sigma]$.

As before, we note that an analogous result holds between predicative fragments of SOL, where the comprehension schema is restricted to formulae not containing bound predicate variables (and possibly no free variables), and subsystems of $\mathsf{UTB}[\sigma]$ in which the instances of disquotation are restricted to the corresponding translations.

Again, we do not recommend $\mathsf{UTB}[\sigma]$ as our ultimate truth theory, but our results show that second-order reasoning can be simulated in a first-order framework. In particular, if Γ is a theory formulated in $\mathcal{L}_2$, then ψ is a theorem of Γ if and only if $\sigma(\psi)$ is a theorem of $\sigma(\Gamma)$, the set of all translations of non-logical axioms of Γ. This implies that our truth theory $\mathsf{UTB}[\sigma]$ is consistent.

Moreover, our method can be extended to show that it is possible to mimic n-th-order logic in a disquotational theory of truth, for every natural number n. One can extend $\mathcal{L}_2$ to $\mathcal{L}_3$ by adding new predicate symbols that may be applied to the predicate symbols and individual terms of $\mathcal{L}_2$. Third-order logic extends second-order logic by a comprehension axiom schema in which the relevant formula φ may now contain third-order quantifiers as well. We can move from $\mathcal{L}_3$ to a language $\mathcal{L}_4$ by admitting further predicate symbols that may be applied to the predicate symbols of $\mathcal{L}_3$, and add appropriate comprehension axioms. In this way, we can construct $\mathcal{L}_n$ and formulate n-th-order logic for every natural number n. Using techniques similar to those discussed before, one can show that for each n, one can find a suitable disquotational theory of truth in which n-th-order reasoning can be carried out. Hence, the function deflationists ascribe to the truth predicate should not be limited to the simulation of sentential quantification in a first-order setting but should also extended to quantification into predicate position in general, for predicates of arbitrary (finite) order.

5. The Conservativity Debate

One of the claims attributed to deflationism and often considered fundamental to the view is that truth is a 'thin', 'lightweight', or 'insubstantial' notion. Many have taken this 'lightness' to mean that the concept of truth does not contribute anything to our knowledge of the world and has no explanatory power. Thus, the insubstantiality of truth has often been wedded to the idea that adding a deflationary truth predicate to a given theory must always lead to a *conservative* extension of it.

The notion of conservativity is one that will be familiar to readers of Hartry Field's book *Science without Numbers*, and has received some attention in recent

discussions about metaontological deflationism (e.g. Schiffer, Thomasson, Hale and Wright). In the present context, it can be defined as follows:

> **Definition 3** (Conservativity). A theory of truth Σ formulated in the language $\mathcal{L}_{\mathrm{T}}$ is conservative over its base theory Γ formulated in $\mathcal{L}$ if and only if all consequences of Σ that do not contain the truth predicate are consequences of Γ as well.

In other words, Σ is conservative over Γ just in case any truth statable in the language of $\mathcal{L}$ that can be derived from Σ must already be available in Γ.

That the lightness of truth should be cashed out in terms of conservativity has been suggested by Horsten [17], Shapiro [30], and Ketland [21]. For example, Ketland [21, p. 79] writes that 'if truth is non-substantial – as deflationists claim – *then* the theory of truth *should* be conservative. Roughly: *non-substantiality* ≡ *conservativeness*.' Shapiro presents the following argument for the conservativity requirement:

> I submit that in one form or another, conservativeness is essential to deflationism. Suppose, for example, that Karl correctly holds a theory B in a language that cannot express truth. He adds a truth predicate to the language and extends B to a theory B' using only axioms essential to truth. Assume that B' is not conservative over B. Then there is a sentence φ in the original language (so that φ does not contain the truth predicate) such that φ is a consequence of B' but not a consequence of B. That is, it is logically possible for the axioms of B to be true and yet φ false, but it is not logically possible for the axioms of B' to be true and φ false. This undermines the central deflationist theme that truth is insubstantial. Before Karl moved to B', $\neg\varphi$ was possible. The move from B to B' added semantic content sufficient to rule out the falsity of φ. But by hypothesis, all that was added in B' were principles essential to truth. Thus, those principles have substantial semantic content. [30, p. 498]

Admittedly, the conservativity requirement has some intuitive force. However, it generates a conflict because most prominent formal theories of truth on the market are *not* conservative over their respective base theories. For example, consider the theory CT (for 'compositional truth'). Roughly speaking, this theory is obtained by turning the clauses of Tarski's semantic definition of truth into axioms and adding them to PA with induction for all formulas of the language, including those containing the truth predicate. For instance, one of CT's truth-theoretic axioms is 'For all sentences x of $\mathcal{L}$, the negation of x is true if and only if x is not true'. PA, the base theory, serves as well as syntax theory, as is usual in the literature.[7]

CT is considered by many as a philosophically sound truth theory. But the theory is not conservative over its base theory. The reason is quite simple. Once we have a well-behaved truth predicate on board, we can prove that all the axioms of PA are true, and that the logical inferences preserve truth. Hence we can show that the base theory must be true, which implies that it is consistent,

a statement that can be expressed in the language of PA. But by Gödel's second incompleteness theorem, PA cannot prove its own consistency, provided it is indeed consistent. Thus, CT allows us to prove a statement in the base language that is not provable in the base theory. (Indeed, Shapiro [30, p. 499] suggests that any adequate theory of truth will allow us to carry out this argument; otherwise it could hardly be called 'adequate'.) Therefore, CT counts as a substantial theory of truth according to those who equate conservativity with insubstantiality.

Some philosophers might be willing to bite the bullet in this case, arguing that CT is not a disquotational but compositional theory truth, and that compositional axioms for truth are not compatible with deflationism anyways. We do not agree with this assessment, but it is worth pointing out that one can also find examples of disquotational theories of truth that are not conservative. One of them is the theory PUTB of positive uniform disquotation (cf. Halbach [13]). It contains an instance of the uniform T-schema for every formula in which the truth predicate occurs only positively, i.e. for every formula that does not contain negated occurrences of the truth predicate.[8]

Unsurprisingly, the conservativity debate has spanned a literature of its own in recent years.[9] For example, Field [6] does not reject the conservativity requirement but claims that CT is conservative if properly understood. In order to prove the consistency of PA in CT one needs to rely not only on the compositional axioms for truth but also on the principle of induction (for formulas containing the truth predicate). Without them, the theory would conservatively extend PA. According to Field, the relevant instances of induction are not purely truth-theoretic, but rather mathematical principles.

This defence has been regarded by many as problematic. On the one hand, it might be difficult to draw a clear line between truth-theoretic and mathematical content (cf. Horsten [18], Halbach [14, pp. 315-316]). On the other, it has been suggested that, if the relevant instances of induction are taken to be mathematical, they should be considered as part of the base theory. In that case, adding the compositional truth principles yields again a non-conservative extension of the base theory.[10]

A different solution was entertained (but ultimately rejected) by Shapiro [31]. He suggests that the notion of consequence employed in the definition of conservativity could be reconceived as second-order logical consequence with so-called standard semantics, where in a model second-order quantifiers are understood as ranging over the full power set of the domain. Now suppose that our base theory is second-order arithmetic. Since this theory is categorical, that is, has only one model up to isomorphism, every true arithmetical sentence is a second-order consequence of the base theory. So there is no way that the truth theory could prove an arithmetical sentence that is not already a second-order consequence of the base theory. Hence the conservativity requirement is satisfied.

Many philosophers have taken issue with this proposal. First, one may be at unease about the notion of second-order logical consequence, e.g. for Quinean reasons: second-order logic comes with 'staggering existence assumptions'.

Second, one may have worries about the epistemic tractability of the relation of second-order logical consequence: it is not effective, and the validities of second-order logic are not recursively enumerable. Third, even if one deems second-order logical consequence acceptable, one may doubt whether it is compatible the with deflationary account of truth (cf. Shapiro [30] and Murzi & Rossi [24]).

We do not share any of these worries; nevertheless, we reject Shapiro's proposal. A minor issue is that even if one is ready to embrace Shapiro's solution in the case where a truth theory is added to arithmetic, it is not obvious that it will work if we add a truth predicate to some other base theory. For assume that Γ is a second-order theory that is not categorical. Then there are sentences that are undecided by Γ, i.e. there are sentences such that neither the sentence nor its negation are implied by Γ. Thus there is a theoretical possibility that combining Γ with a truth theory leads to a non-conservative extension. One may think that it is unlikely that a theory of truth will decide a sentence that is undecidable in Γ (given that Γ is closed under second-order logical consequence), but as McGee [23] has shown, any sentence whatsoever can in principle be decided by an instance of the T-schema.

What really concerns us though is that one reads so much into the insubstantiality of truth that it forces the deflationist to appeal to a notion (i.e. second-order consequence) that has sparked so much controversy. Whether the notion of second-order consequence is in good standing or not is a difficult question: embracing a deflationary view on truth should not force one to adopt a particular position on this issue.

Waxman [34] has proposed an interesting disjunctive strategy. Our conception of arithmetic is either a categorical one or an axiomatic one, based e.g. on PA. We need not take a stand on which conception is the right one, but whichever it is, Waxman argues, we will end up with a truth theory that satisfies the conservativity requirement. Assume first our conception of arithmetic is the categorical one. Then we must be in possession of resources that allow us to rule out all non-standard models. Whatever these resources are, they will induce a notion of consequence that goes beyond the proof-theoretic one, and we should adopt that notion in our definition of the conservativity requirement. As we have seen above, our truth theory will not conflict with the conservativity requirement if the latter is defined that way. Now assume that our conception of arithmetic is a non-categorical, axiomatic one instead. Then we have a principled reason to rule out any truth theory that proves arithmetical sentences that are not entailed by our axiomatic theory of arithmetic (say, PA). Thus, if our conception of arithmetic is the axiomatic one, we should adopt a proof-theoretic notion of conservativity and adopt a theory of truth that satisfies the constraint. Either way, we will end up with a truth theory that meets the conservativity requirement.

We are not satisfied with the disjunctive strategy. To begin with, the disjunctive strategy seems somehow to presuppose (or only works if we presuppose) that we either have a categorical conception of arithmetic or one based on some particular axiomatic theory. But couldn't it be that we have some 'open-ended'

understanding of arithmetic which allows us to endorse stronger and stronger axiom systems as we go along without, however, ever being in a position to rule out each and every non-standard model? If this were the case, we could neither appeal to the notion of second-order consequence in the definition of the conservativity requirement, nor would we be able to say of some truth theory in advance whether it meets the proof-theoretic conservativity requirement or not: a truth theory might seem unacceptable because it proves sentences that are not provable in our current system of arithmetic, but later on we discover that it was acceptable after all because meanwhile we have adopted a stronger system of arithmetic (or perhaps the status of the truth theory switches from unacceptable to acceptable in the moment we adopt the stronger system).

Even if we put this possibility aside, the disjunctive strategy cannot tell us whether CT, say, is an acceptable theory of truth or not. It would seem that we first need to find out what our conception of arithmetic is. Do we all share the same conception of arithmetic, and is it an empirical question which one it is? Or does everyone have their own conception of arithmetic, depending on one's philosophical inclination? If the former is the case, should we all abstain from using a theory like CT until we have found what our common conception is? If the latter is the case, should everyone just pick their own truth theory, in conformity with the conservativity requirement that follows from their own conception of arithmetic?

Moreover, we suspect that what the proponents of the conservativity requirement have in mind is that a truth theory is deflationary as long as it is conservative, not over logic, for that is impossible, but over *any* base theory containing enough syntax, as discussed by Halbach [12]. Regardless of our epistemic preferences, a deflationary truth predicate, they would maintain, should not *inform* us of any new facts about the ontology of the base theory. We therefore think that the disjunctive strategy does not rise to the conservativity challenge.

It turns out there is a much easier way out for the deflationist. Indeed, we would like to suggest that conservativity is not a reasonable requirement to begin with, and its association with (the version of) deflationism (we laid out in section 1) is the consequence of a misinterpretation of the insubstantiality claim.

Assume, for example, that one wishes to generalise into predicate position in the context of PA. The most popular option is to work within the framework of second-order logic. Once second-order quantifiers are available, it's fair to exploit their expressive power to provide finitary reformulations of the axiom-schemata of the original theory; in this case, to replace the Induction schema with the Induction axiom. As a result, we obtain the theory Z_2. As is well known, Z_2 is a non-conservative extension of PA, for it proves PA's consistency statement expressed in the language of PA itself. It does so in virtue of the principles and rules of inference added for the second-order quantifiers, that increase the deductive power of the underlying logic, but also due to the fact that some

axioms of the original theory were reformulated in terms of these new logical operators.

We have argued at length that the role deflationists ascribe to the truth predicate should be taken to include the ability to mimic second-order quantification. If we are right, we should be able to find a disquotational truth predicate in terms of which we can provide finitary reformulations of the axioms of PA and that allows us to draw inferences from these new axioms, just as in the case of Z_2. If the job is properly done, it is only to be expected that the reformulated arithmetical principles in our truth theory entail sentences not containing the truth predicate that were not provable already in PA, just as Z_2 does. Indeed, this is exactly what Theorem 2 shows. The translation, σ, maps formulas of PA into truth-free expressions. Thus, in $\mathsf{UTB}[\sigma]$ $\sigma(Z_2)$ will entail the translation of PA's consistency statement, which is not provable from $\sigma(\mathsf{PA})$ itself. In other words, reformulating arithmetic availing of our disquotational truth predicate does not yield a conservative extension of the original theory. This should be neither disappointing nor surprising, but actually desirable. The adoption of second-order resources occasionally results in non-conservative extensions of first-order theories. If a disquotational truth predicate allows us to simulate those resources, as we maintain, then disquotational truth theories will occasionally result in non-conservative extensions of first-order theories as well. Not only is deflationism untied to the conservativity requirement, but also deflationist should in certain cases actively seek non-conservative truth theories.

Is our analysis of the function of truth incompatible with the insubstantiality claim, so pervasive among deflationists? We don't think so. If it were, there would either be a serious flaw in our proposal or in the deflationist position as a whole. However, as we indicated a few paragraphs above, the association of the insubstantiality claim with the conservativity requirement is a product of a misguided interpretation of the former, which should be understood instead in the context it originated. In what follows we examine the three versions of the insubstantiality claim we could identify in the deflationist literature and argue that none of them really entails the conservativity requirement.

Proto-versions of the insubstantiality claim can be traced back to redundantism. In light of the equivalence thesis, many exponents of the view, including Frege [7], Ramsey [28], Ayer [1], and Strawson [32], saw in the truth predicate a rhetoric device rather than an expression standing for a legitimate property – at least according to the sparse conception of properties. For instance, Frege [7, p. 293] wonders, 'may we not be dealing here with something which cannot, in the ordinary sense, be called a quality at all?' Following some sort of inference to the best explanation, redundantists rejected the need and possibility of finding a real definition that would unveil the *metaphysical nature* of truth, for there is none. At least not one that could be described in terms of correspondence, coherence, serviceability; no essence that would make truth a concept in the traditional sense. As Ayer puts it,

> the words 'true' and 'false' are not used to stand for anything, but function in the sentence merely as assertion and negation signs. That is to say, *truth* and *falsehood* are not genuine concepts. Consequently, there can be no logical problem concerning the nature of truth". ([1, p. 28-29])

Contemporary versions of this idea can be found, e.g. in Horwich [20], according to whom truth is a property only in the abundant sense, as it lacks a constitution.

We hardly see how one could infer the conservativity requirement from this version of the lightness claim. A predicate merely satisfying the equivalence between each sentence and its truth ascription might be just an expressive device that doesn't express a (real) property. But, as our results show, its expressive powers are nonetheless strong enough to facilitate, in some cases, the proof of statements in the language of the base theory that the latter does not entail.

A second trace of the lightness claim that builds on the former can be found already within deflationism properly understood. It emanates from the position's second core thesis, namely, that the truth predicate just plays an expressive role. Although already implicit in Quine [27] (see section 2 above), modern deflationists often maintain that the truth predicate is akin to a logical device and truth to a logical property (cf. Horwich [20], Field [6], Horsten [18]). As such, truth has no deeper metaphysical nature than, say, conjunction. Just like conjunction, it is entirely truth-functional, for disquotational principles in the object language – or, equivalently, corresponding semantic clauses in the metalanguage – suffice to account for every role the truth predicate can play. This means, again, that there is no room for a more intensional account of truth that would disclose a thick essence. Actually, deflationism takes its name from this idea of deflating the sort of metaphysical inquiry of truth pursued by correspondentists, coherentists, etc. already present in the era of redundantism and bolstered by the assimilation of the second core thesis.

Does assigning truth a quasi-logical status commit us to the conservativity of our truth principles over their respective base theories? We believe this is not the case. On the one hand, it is not even clear whether logical notions are conservative. For example, consider the fragment of propositional intuitionistic logic that only contains $\wedge$, $\vee$, and $\rightarrow$. The addition of the rules for the Boolean negation $\neg$ to this logic does not yield a conservative extension of it, for the extended logic proves Peirce's law: $((\varphi \rightarrow \psi) \rightarrow \varphi) \rightarrow \varphi$, whereas the original fragment does not. If classical negation is to be counted among the logical expressions, conservativity cannot be a general requisite for logicality. However, some authors do maintain that conservativity should be required, ruling out classical negation as logical. From their point of view, if a truth predicate were to be logical, it should be conservative as well.

But on the other hand, deflationists don't claim that truth is logical but only that it's *quasi*-logical. For instance, Horsten maintains that 'the notion of truth is wrapped up with syntax, which is wrapped up with finite combinatorics,

which finally reduces to arithmetic. Instead of being a purely logical notion, truth should at least be called a *logico-linguistic* or *logico-mathematical* notion.' [19, p. 85] Given our analysis of the function of the truth predicate, namely that it enables us to simulate higher-order quantification within a first-order framework, we would like to suggest the following:

> *The truth predicate has the same logical status as sentential and predicate quantifiers.*

We need not decide here whether sentential and predicate quantifiers are indeed logical. This is perhaps just a matter of terminology. The point here simply is: the truth predicate and sentential/predicate quantifiers fall into the same category, they both serve the same purpose. As we mentioned already, second-order logic certainly does not always yield conservative extensions over some sets of premises, e.g. Z_2 is not conservative over PA. So, again, if a truth predicate can serve as a device for mimicking higher-order quantification, as we submit, then it's neither bad nor puzzling that the truth systems that are up to the task are not conservative over their respective base theories. It is to be expected.

Finally, explicit adherence to the insubstantiality claim can be found in modern deflationism, also linked to the second core thesis of the view. It has been argued that, if its only function is the logico-expressive one examined in section 2, the truth predicate cannot play a central role, e.g. in explaining the nature of meaning, knowledge, or the goals of belief and assertion, as is widely believed. At least not a much more substantive role than that of the logical operators that occur in such explanations. In Horwich's [20, p. 4] words, 'It is just in this role, and not as the name of some baffling ingredient of nature, that the concept of truth figures so pervasively in philosophical reflection.' Deflationists often encapsulate this idea under the slogan that truth has no explanatory power.

This is perhaps the version of the lightness claim that most strongly suggests conservativity should be required of deflationary truth predicates. The argument typically goes along the following lines. Proofs of new facts about a certain portion of reality are usually taken to have explanatory value. Thus, a deflationary truth predicate is not allowed to play an essential role in such proofs. In other words, the addition of deflationary truth principles to a base theory cannot entail more statements in the language of the base theory than the base theory itself; it must be conservative.

However, 'no explanatory role' is just a slogan. What was originally meant by it, as we pointed out above, is that the only role the truth predicate can play in any sort of explanation is the one discussed in section 2. To repeat, this is essentially to mimic higher-order quantification. Since the latter occasionally features in proofs of new facts about the ontology of the base theory, so does the truth predicate; these cases do not contradict the version of the insubstantiality claim under discussion. Furthermore, if deflationism is right and this is

the only function of truth, in no circumstance will the truth predicate play a substantive role. As Field [6, p. 537] puts it: 'any use of "true" in explanations which derives solely from its role as a device of generalization should be perfectly acceptable'.

6. Conclusion

We have thoroughly examined the function deflationism attributes to the truth predicate and argued that it is best understood as simulating higher-order quantification in a first-order setting. Our analysis has many important consequences for our understanding of deflationism. In the last section we have touched upon the conservativity requirement for deflationary truth theories and argued that it's inappropriate. There are many more consequences. They will be dealt with in subsequent articles. However, let us close this paper by at least briefly mentioning some of them.

Properties. Some authors take second-order quantifiers to range over properties, and understand second-order logic as a theory of properties (e.g. Hale [15]). Above we have seen that second-order quantification is reducible to a truth theory. This seems to suggest that property talk could also be a means for quantifying into the predicate position and, accordingly, leads to a deflationary understanding of property talk as well.

The lightness of truth. A slogan that is often associated with deflationism is that 'truth is simple'. However, this should be understood as referring either to the predicate or to the concept of truth, but not to the truths themselves; let alone to the procedures for finding out whether a given statement is true. Deflationism is in no way committed to the view that the extension of 'is true' is of low computational complexity.

Classical v non-classical logics. Many philosophers have argued that deflationists are committed to the unrestricted T-schema and therefore to a non-classical logic, in order to block the paradoxes. Based on our analysis, one could argue that deflationists are not committed to any particular logic. We have seen that classical truth theories can give us the full expressive power of the simple theory of types. A non-classical truth theory may allow us to generalise over even more sentence/predicates in the language. On the other hand, adopting a non-classical logic means that we have to reject some meaning postulates for other logical devices such as negation or the conditional, which is likely to result in a loss of proof-theoretic power overall. Therefore, all other things being equal, deflationism motivates at best a logical pluralism of sorts, rather than a ban on classical logic.

Semantic vs axiomatic approaches. It is often argued in the literature that deflationism is committed to some form of inferentialism, as the truth predicate should be introduced axiomatically. If semantic clauses are employed instead, the argument goes, the truth theory commits one to non-deflationary,

substantial truth principles. At the bottom of this kind of reasoning is the idea that Tarskian compositional definitions of truth and satisfaction put forward substantial notions. Based on our analysis of the function, and thus the essence, that deflationists attribute to truth and satisfaction, we will argue to the contrary.[11]

Notes

1. For more details on this, see Heck [16].
2. Moreover, the fact that we allow operators in $\mathcal{L}$ makes our relative interpretability result below non-trivial. In the absence of non-logical operators, sentential quantifiers can be defined in first-order logic: simply let $\forall\alpha_i\,\varphi$ be short for $\varphi[\top/\alpha_i] \wedge \varphi[\bot/\alpha_i]$, where $\top$ is a any logical truth and $\bot$ any logical falsity. However, this proof cannot be generalised to theories whose language contains additional operators. For instance, consider the system S5 formulated in QPL (cf. Kripke [22]), where $\Box$ stands for necessity. If φ is a sentence of the language that is neither necessary nor impossible, then $\exists\alpha\,\Box(\alpha \leftrightarrow \varphi)$ is not provably equivalent to $\Box(\top \leftrightarrow \varphi) \vee \Box(\bot \leftrightarrow \varphi)$.
3. One could consider extending the language with a term-forming operator $\ulcorner . \urcorner$, such that if φ is a formula, $\ulcorner\varphi\urcorner$ is a name for φ, i.e. a term. However, this move is highly controversial, as it would jeopardise the consistency of the system. See, for instance, Tarski [33, §1]. Alternatively, Azzouni [2] has developed an interesting (although predicative) system in which one can simultaneously bind sentential and nominal places. We believe that our results can be extended to his system as well, i.e. there is a natural translation from his system into a disquotational theory of truth.
4. It might seem that the definition of τ is circular because in the translation of certain formulas we need to rely on the predicate $\mathrm{Trsl}(x)$ that applies exactly to those formulas that are in the range of τ. However, the existence of such a translation is guaranteed by Kleene's Recursion Theorem.
5. The translation of negated formulas are slightly different from what one would normally expect. The straightforward procedure would be to translate $\neg\varphi$ as $\neg\tau(\varphi)$. This is not possible because we did not translate negated atomic formulas in this way. For example, we translate α as $\mathrm{Trsl}(x) \rightarrow \mathrm{T}x$ but we translate $\neg\alpha$ as $\mathrm{Trsl}(x) \rightarrow \neg\mathrm{T}x$ rather than $\neg(\mathrm{Trsl}(x) \rightarrow \mathrm{T}x)$.
6. For more on second-order logic, see Shapiro [29].
7. For further details, see e.g. Horsten [19, chap. 6] or Halbach [14, chap. 8].
8. An essential ingredient of the liar paradox is that it contains a negated occurrence of the truth predicate. The idea behind PUTB is to avoid all paradoxes by banning sentences with negated occurrences of the truth predicate. The theory is indeed consistent in classical logic and proof-theoretically very strong: it interprets the non-conservative Kripke-Feferman theory KF.
9. For an overview and more references, we send the reader to Cieslinski [5].
10. This needs some reformulation of the definition of conservativity because now we have allowed the truth predicate to occur in the language of the base theory (cf. Horsten [19, p. 80]).

11. This paper was written while the second author held an ERC Marie Curie Fellowship (792202, LOFUPRO).

References

1. Ayer, A. J. The Criterion of Truth. *Analysis* 3 (1935), 28–32.
2. Azzouni, J. Truth via anaphorically unrestricted quantifiers. *Journal of Philosphical Logic* 30 (2001), 329–354.
3. Azzouni, J. Anaphorically unrestricted quantifiers and paradoxes. In *Deflationism and Paradox*, J. C. Beall and B. Armour-Garb, Eds. Oxford University Press, 2005, pp. 250–273.
4. Carnap, R. *Introduction to Semantics*. Harvard University Press, 1948.
5. Cieśliński, C. *The Epistemic Lightness of Truth. Deflationism and its Logic*. Cambridge University Press, Cambridge, 2017.
6. Field, H. Deflating the Conservativeness Argument. *Journal of Philosophy* 96 (1999), 533–540.
7. Frege, G. The thought: a logical inquiry. *Mind* 65 (1956), 289–311.
8. Grover, D. L. Propositional Quantifiers. *Journal of Philosophical Logic* 1 (1972), 111–136.
9. Grover, D. L., Camp, J. L., and Belnap, N. D. A Prosentential Theory of Truth. *Philosophical Studies: An International Journal for Philosophy in the AnalyticTradition* 27 (1975), 73–125.
10. Gupta, A., and Belnap, N. D. *The Revision Theory of Truth*. MIT Press, Cambridge, 1993.
11. Halbach, V. Disquotationalism and infinite conjunctions. *Mind* 108 (1999), 1–22.
12. Halbach, V. How innocent is deflationism? *Synthese* 126 (2001), 167–194.
13. Halbach, V. Reducing compositional to disquotational truth. *Review of Symbolic Logic* 2 (2009), 786–798.
14. Halbach, V. *Axiomatic Theories of Truth*. Cambridge University Press, Cambridge, 2011.
15. Hale, B. Properties and the interpretation of second-order logic. *Philosophia Mathematica* (2012), 1–21.
16. Heck, Jr, R. Truth and Disquotation. *Synthese* 142 (2004), 317–352.
17. Horsten, L. The semantical paradoxes, the neutrality of truth and the neutrality of the minimalist theory of truth. In *The many problems of realism*, P. Cortois, Ed., vol. 3 of *Studies in the general philosophy of science*. Tilburg University Press, 1995, pp. 173–187.
18. Horsten, L. Levity. *Mind* 118 (2009), 555–581.
19. Horsten, L. *The Tarskian Turn: Deflationism and Axiomatic Truth*. MIT Press, Cambridge, 2011.
20. Horwich, P. *Truth*, second ed. Oxford University Press, 1998.
21. Ketland, J. Deflationism and Tarski's paradise. *Mind* 108 (1999), 69–94.
22. Kripke, S. A Completeness Theorem in Modal Logic. *Journal of Symbolic Logic* 24 (1959), 1–14.
23. McGee, V. Maximal consistent sets of instances of Tarski's schema. *Journal of Philosphical Logic* 21 (1992), 235–241.
24. Murzi, J., and Rossi, L. Conservative Deflationism? *Philosophical Studies* (2018).
25. Picollo, L., and Schindler, T. Disquotation and infinite conjunctions. *Erkenntnis* (2017), https://doi.org/10.1007/s10670-017-9919-x.
26. Putnam, H. Meaning and knowledge. In *Meaning and the moral sciences*. Routledge, London, 1978, pp. 1–80.
27. Quine, W. V. O. *Philosophy of Logic*. Harvard University Press, 1970.
28. Ramsey, F. P. Facts and propositions. *Proceedings of the Aristotelian Society* 7 (1927), 153–170.
29. Shapiro, S. *Foundations without Foundationalism: A Case for Second-Order Logic*. Oxford University Press, New York, 1991.

30. Shapiro, S. Proof and Truth: Through Thick and Thin. *Journal of Philosophy* 95 (1998), 493–521.
31. Shapiro, S. Deflation and conservation. In *Principles of Truth*, V. Halbach and L. Horsten, Eds. Ontos, Frankfurt am Main, 2002, pp. 103–128.
32. Strawson, P. F. Truth. *Analysis* 9 (1949), 83–97.
33. Tarski, A. The concept of truth in formalized languages. In *Logic, Semantics, Metamathematics*. Clarendon Press, Oxford, 1935, pp. 152—278.
34. Waxman, D. Deflationism, Arithmetic, and the Argument from Conservativeness. *Mind* 126 (2017), 429–463.