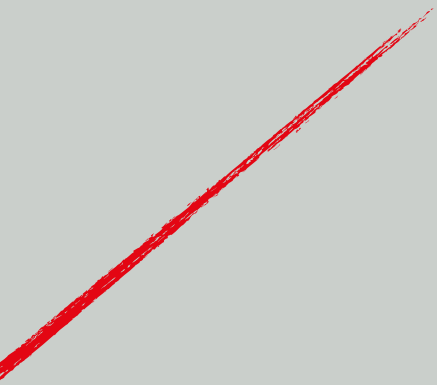
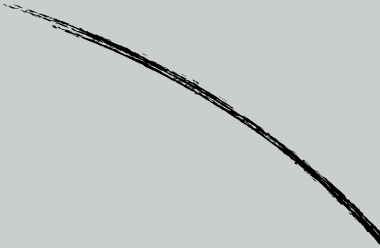
RATIONES

Andrea Strollo

The mathematics of deflationary truth





PADOVA UP

PADOVA UNIVERSITY PRESS

Rationes è una collana filosofica open access che ospita testi originali sottoposti a *double blind peer review*.

Rationes

Volume realizzato con il contributo di
Progetto CARIPARO: Polarization of irrational collective beliefs in post-truth societies. How anti-scientific opinions resist expert advice, with an analysis of the anti-vaccination campaign (PolPost).

First edition 2021 Padova University Press
Original title *THE MATHEMATICS OF DEFLATIONARY TRUTH*

Università degli Studi di Padova
via 8 Febbraio 2, Padova
www.padovauniversitypress.it

Editing
Padova University Press

ISBN 978-88-6938-241-3

Andrea Strollo

THE MATHEMATICS OF DEFLATIONARY TRUTH

PADOVA UP

Metaphysica sunt, non leguntur.
Mathematica sunt, non leguntur.
(G. Frege)

Table of Contents

PREFACE

It is usually hard to tell what makes a scholar interested in something rather than something else, but truth is one of the topics I have thought about the most in the last years. Apart from sociological reasons, of which I might not be completely aware, there are some theoretical motivations. These motivations mostly have to do with the fact that truth is at the juncture of human representation and reality. It is where language and world connect. As such, it is a key notion to disentangle these two sides, straightforwardly connecting it with huge philosophical issues, such as the analytic/synthetic divide, the distinction between semantics and metaphysics, and so on. Investigating the notion of truth was then a natural consequence of my personal attitude to address fundamental issues and go to their core. Since I am not an exceptional human being, I expect such an attitude to be shared by others, who might find the topic equally engaging.

But how to study truth? Here comes the second autobiographical aspect. Apart from Philosophy, I have always been interested in Logic. The application of formal methods, so typical in Analytic Philosophy, was the natural choice again. Of course, also in this case, theoretical motivations are available. In particular, I believe that theoretical progress requires a huge effort to gain clarity, and the application of formal tools is one of the best ways to achieve that. I hope that this work contributes to prove the

fruitfulness of such an approach by offering a concrete case of this progress.

This book is mostly based on my doctoral thesis, which I have extended and updated. I wrote the dissertation under the guide of Diego Marconi, and defended at the University of Turin in front of a committee composed of Andrea Cantini, Volker Halbach, and Gabriele Usberti. I mention this because I am particularly proud of having received my PhD from them, and also because I take it to show the connection with the Italian tradition in Logic and Analytic Philosophy, under which this work was developed. The title is an implicit homage to a great Italian logician, Ettore Casari, who recently passed away. The title of this book is intended to echo his *La Matematica della Verità*.

Many other people, who made this book possible for various reasons, should be acknowledged. In particular, I want to mention and thank Massimiliano Carrara, Massimo Mugnai, Tianqun Pan, and Gabriel Sandu, who believed in me and my work. If anything good is in this book, it is dedicated to them, and to all those who supported me during these years. Since books inevitably also contain mistakes or imprecisions, I want to dedicate all such neglected errors to those who made my studying, working in academia, and living in general, more difficult and limited than it could have been. Responsibility for the mistakes is ideally shared with them.

This work has been funded by Progetto CARIPARO: *Polarization of irrational collective beliefs in post-truth societies. How anti-scientific opinions resist expert advice, with an analysis of the anti-vaccination campaign (PolPost).*

INTRODUCTION

"What is truth?" Pilate's question is one of the most typical and profound in philosophy. Unsurprisingly, tentative answers abound. This is particularly the case for analytic philosophy since, in it, the problem of truth has been the object of systematic investigations like it never was. Such attention has kept increasing in the last decades. The reason is quite clear. Apart from being a traditional philosophical topic, representing a crucial juncture where representation and reality merge, truth plays a fundamental role in disciplines that characterize the core of analytic philosophy: logic and semantics. Joining together the philosophical and the logical side is also one of the marks of this book. Under this respect, the present work fully aligns with the analytic tradition. Indeed, it also has the ambition to offer an implicit defence of such an approach, showing how much can be gained by keeping philosophical depth and logical precision together. This book is written with the conviction that progress in philosophy is possible, and that one of the best ways to realize it is to promote a strict collaboration between logic and philosophy. Whether I successfully achieved this aim is left to the reader to decide.

Analytic conceptions of truth can be broadly classified along two lines, depending on how they answer the following two questions: 1. Has truth a substantial nature? 2. How many properties of truth are there? Traditional views, like correspondence, coherence, pragmatist views

all answer "yes" to both questions. The second half of the last century and the last decades in particular, however, has witnessed the rapid growth in popularity of conceptions of truth answering "no" to at least one. Pluralist views hold that there is more than one property of truth, whereas deflationary views hold that truth lacks a substantial nature. Pluralist and deflationary views are today the main rivals in the field. This book focuses on deflationary views of truth and contributes to the general contemporary debate on truth by addressing an issue emerging from an appealing way to clarify what the insubstantiality of (deflationary) truth might amount to: conservativeness.

At the end of the nineties some authors (Leon Horsten, Stewart Shapiro and Jeffrey Ketland) put forward a fairly technical argument against deflationary theories of truth. In a nutshell, deflationism, it was argued, is committed to conservativeness by the the claim that truth is not a substantial notion, however a conservative theory (under the light of certain logico-mathematical facts) cannot arguably be an adequate theory of truth, so that deflationism is an inadequate theory of truth. Despite its apparent simplicity, the argument involves several subtle issues on which deflationists are forced to take a stand. Deflationists reacted in different ways, but the general final impression is that deflationism does seem committed to conservativeness and is condemned to be, if not an utterly inadequate conception, at least a very weak one. In this book I enquiry into this assessing how and in what measure such a received outcome is correct.

This work is divided into three main parts. The first part is a very general survey where the philosophy and mathematics of truth is introduced. In chapter one we sketch philosophical conceptions of truth comparing more traditional approaches to deflationism; we briefly survey the history of deflationism and a general characterization

is put forward. In the second chapter, the notion of truth is considered from a formal perspective and some fundamental axiomatic theories of truth are introduced.

The second part joins the philosophical and the formal approach to truth together into the debate over deflationism and conservativeness. In chapter three we introduce the notion of conservativeness and we give some examples of its applications both to logical and philosophical issues. Then we spell the argument from conservativeness out and discuss it in order to extract a precise requirement that deflationary theories are supposed to satisfy. In chapter four we focus on deflationist replies to the argument from conservativeness and we critically discuss each solution.

The third and last part consists in a more technical study of some presuppositions of the debate: we want to take a step back and compare the two major claims of deflationism - the centrality of T-sentences and the logical function of the truth predicate - with conservativeness. Although most of the results discussed there are available in literature, the originality consists in gathering them together and systematically reviewing them under the light of the conservativeness argument. More original reflections can be found in chapter seven and in chapter eight. In Chapter Five we compare T-sentences with the empty base theory and in Chapter Six we analyse in what measure a deflationary theory can be really conservative over a theory of syntax. The result would be quite serious for a deflationist. In Chapter Seven we compare conservativeness with the logical function of the truth predicate. We will get the unpleasant result that the truth predicate is not able to serve the logical function in no sense without losing conservativeness at the same time. In Chapter Eight we draw some conclusions and sketch a reformulation of the conservativeness requirement. This new requirement, we will argue, does justice to the insubstantiality of truth and at the same time it does not condemn deflationism to death.

PART ONE

CHAPTER ONE
DEFLATIONISM AND ITS RIVALS[1]

Truth plays a central role in virtually every theoretical activity and is essential both in our ordinary life and in science. It is usually taken to be the goal of scientific inquiries and a valuable property of our beliefs. Besides its general relevance, truth is also important in several particular disciplines.We need it in epistemology, for instance, since, according to the classical view, knowledge is justified *true* belief. We need it in philosophy of science, to properly articulate scientific realism in its semantic dimension. We need it in metaphysics to understand, beside the realist/antirealist divide, truth making and grounding. The relevance of the notion in semantics is even more upfront. The traditional approach to meaning, in fact, is explicitly built on analysis of truth-conditions. Nowadays we have important lines of research departing from this standard approach (like cognitive semantics, to name one), but truth-conditional semantics remains the most developed and possibly dominant one. Whether we want to work in that paradigm or to criticize it, it is worth having an answer to the question of what a truth-condition is supposed to be. Clearly, truth is also fundamental in logic, since logical consequence is traditionally understood in terms of necessary truth preservation.

[1] For general references see Kirkham 1992, Künne 2003.

It is not hard to see why truth is such an important notion, since it is a central point of contact of thought, language, belief, knowledge, reality, action, etc. to understand it is to disentangle and clarify some of the deepest aspects of reality and inquiry. Hence, an answer to Pilate's question naturally presents itself as one of the most philosophically valuable and challenging at the same time.

SUBSTANTIALIST CONCEPTIONS OF TRUTH

Although many important thinkers in the history of philosophy have offered reflections on truth (with Aristotle, Thomas Aquinas, Ockham, Kant being major examples), explicitly and more developed conceptions have been proposed only in recent times, with the rise of analytic philosophy. During the twentieth century traditional ideas have been especially worked out and new approaches have been explored. Beside more classical views of truth like correspondentism, coherentism or pragmatism, notable innovative approaches have been more recently advanced with the introduction of pluralist and deflationist views.

Analytic conceptions of truth can be broadly classified along two lines, depending on how they answer the following two questions: 1. Has truth a substantial nature? 2. How many properties of truth are there? Traditional views all answer "yes" to both questions. Accordingly, it is assumed that there exists one and only one property of truth, that has some interesting or deep nature of some sort, and that the philosophical task is that of unveiling such a nature offering an account of what truth is. In the attempt to fulfil this task, philosophers endorsing traditional views proposed to understand the nature of truth in terms of correspondence with facts, coherence with some set of beliefs, practical success, verifiability at the end of inquiry

or in ideal circumstances, and so on and so forth. Let us have a closer and quick look at each of these.

Correspondence views develop the venerable idea that truth consists in a particular relation holding between a proposition (or some other kinds of truth bearer) and reality, and this relation is a sort of "correspondence". Truth is correspondence to reality. In this view truth is taken to be a binary relational property involving two kinds of terms: a certain truth bearer and a certain portion of reality. The intuition is intended to align and articulate the commonsensical idea that a proposition is true if it agrees with facts. If it describes the world as it is. The obviousness of this standpoint is the major reason in favour of a correspondence approach and it makes the view the oldest and most traditional one. Note, however, that as long as the idea is not the result of a theoretical elaboration, but only grounded on idiomatic trivialities, it can fit other theories as well. In other words, all conceptions agree that true propositions tell us how things are. The advocates of correspondence, though, consider this claim only as a starting point to elaborate more sophisticated approaches that go fairly beyond this simple intuition. In order to do that, the first task is to clarify what is meant by "correspondence", "proposition" and "reality". Correspondence is supposed to be a relation, apparently a binary relation, but what is this relation like? It involves propositions, or some other truth bearers, but what is the nature of such entities? It also involves a certain portion of reality but, again, what does this exactly mean? In order to solve these problems and to articulate a full fledged correspondence theory we need to engage into a wide philosophical investigation. In correspondentist approaches the construction of a theory of truth easily turns into the construction of an entire philosophical system. For example, a traditional proposal put forward by Wittgenstein has it that correspondence is to be

understood in terms of isomorphism between a proposition and a fact. The view is arrived at by progressive abstraction on the idea of pictorial representation. Although the approach certainly adds many details to the initial intuition, it does so at the cost of raising many other difficulties. For example, according to it a true proposition has in itself a guide to the underlying deep structure of facts. We can then understand reality just by enquiring into language. This looks suspicious to say the least. Another problem, which is familiar to most versions of correspondence views, is that it apparently commits to the admission of a plethora of disparate metaphysical facts. To vindicate the truths of, say, mathematical or moral truths, one should admit mathematical and moral facts as well. Buying the view then easily becomes highly philosophically costly. Being so, it is not surprising that, despite the appealing initial intuition of correspondence, such approaches have not built consensum and are rejected by many philosophers.

Perhaps, we could think, correspondence approaches are wrong in their basic assumption: possibly, truth is not some relation with reality. Instead it is a relation with other truth bearers. If we follow this path, we are led to a coherence theory of truth. The main differences with the previous proposal are two: the relation of correspondence is turned into a coherence one, and the second term of relation is no longer a portion of reality but a set of other propositions or truth bearers. Chief motivations for such an approach can be epistemological or metaphysical. An example of an epistemological motivation is that we can never get outside our beliefs in order to compare one of them with reality: we can at most compare a proposition with other propositions[2]. On the metaphysical side, advocates of this view deny, for instance, that there is a possible distinction between beliefs

[2] Hempel 1935, Neurath 1983, Rescher 1973.

and what makes them true[3]. There is no room for facts, for objects or for the kind of truth makers that correspondence theorists need. A coherence theory thus fits well with idealistic conceptions of reality. Coherence approaches must clearly withstand complications similar to correspondence views. It must be specified what is meant by "coherence", and what set should a proposition be coherent with. They also seem to promote some problematic metaphysics like Idealism. If such a metaphysics is not endorsed, the approach seems hardly able to do justice to propositions true about the external world. Dostoievski's *The brothers Karamazos* is possibly coherent but not true.

Other views alternate to correspondence are pragmatic and epistemological conceptions. In Pierce's version[4] truth coincides, roughly, with what will be held at the end of enquiry: a belief is true if it will be eventually shared by all those who investigate it. His idea moves from the fact that if a person is given enough time she finally will reach a certain view that is the same another person would have reached in her place. Different minds hence tend to agree and their conclusion is true. It is not the case that we all agree on something false. This is not just an optimistic attitude, Pierce takes being true exactly as being what we will finally be agreed on. Truth is the capacity to resist doubt. Epistemological views of truth, like the ones proposed by Putnam and Wright, are somehow similar to Pierce's. In a certain sense Putnam's[5] proposal is an attempt at overcoming some difficulties of the original idea of Pierce. Putnam argues that truth cannot be just rational acceptability; it must be identified with *idealized rational acceptability* instead. We call true a sentence that

[3] Blanshard 1939, Joachim 1906.

[4] Peirce, *Collected Papers*.

[5] Putnam changed opinion about metaphysics and truth some times during his career. Here we consider Putnam's view especially during the phase of so called "internal realism" (Putnam 1981).

would be acceptable under ideal epistemic conditions. The point of seeing truth in terms of epistemic notions, instead of correspondence or the like, is the idea that every truth should be accessible, in principle, to some man in some circumstance or time[6]. Wright further elaborates on this and tries to refine Putnam's view by focusing on the notion of super-assertability, where a proposition is super-assertable if it is assertable in a state of information and remains such in every extension of that state[7]. A big problem of these views is, for example, the difficulty of vindicating truths about inaccessible portions of reality (like the far past, black holes or possible mathematical truths like Goldbach's conjecture).

As emerges from this quick sketch of traditional conceptions, all such views run in serious troubles. As a result, an increasing number of philosophers has recently started to think that the entire debate relied so far on wrong assumptions. Taking for granted claims that are not correct. Accordingly, to solve the riddle of the nature of truth, we should either reject 1. that truth is one, or 2. that truth is something. Truth pluralists follow the first route. They agree with traditional approaches that truth has (at least sometimes) a robust nature to be unveiled, but they deny that there is only one property truth consists in. There are many natures of truth, not just one. One of the main motivations of the view is that traditional conceptions seem to work well in some limited areas of discourse, but they have serious problems when extended to cover all possible areas. For example, a correspondence conception looks appealing when middle size concrete objects are concerned but it is at least puzzling when mathematical or moral claims are considered. By contrast, a coherence view naturally settles mathematical issues, but it can hardly be applied to middle

[6] Notice that it is this claim that Putnam denies in later works.
[7] Wright 1992.

size concrete objects. Rather than trying to squeeze these approaches together, the pluralist proposes to take this fact at face value admitting that the nature of truth can vary in different areas. According to one of the chief ways of articulating the view, truth pluralism holds that there are many properties of truth. Thus, while there may be only one concept and one truth predicate in the language, the expressed properties will typically vary in different areas of discourse. For instance, when mathematics is concerned, truth might be coherence with certain axioms, whereas in talks about history it might consist in correspondence to facts. The basic idea is that truth, or at least its concept/predicate, is characterized by a set of platitudes or truisms that are satisfied by different properties in different areas of discourse. Depending on whether such a set of truisms gives rise also to a generic property of truth on its own, two versions can be obtained: strong and weak. A strong version holds that truth is always area specific and no generic truth is to be admitted, whereas moderate pluralism contends that in addition to a plurality of local truth properties there is also a generic property of being true.

WHY TO BE A DEFLATIONIST

Deflationists disagree both with traditional approaches and with pluralist ones. According to them, the wrong assumption is not that there is a single property of truth, but that such a property has an underlying nature waiting to be discovered. Such an assumption leads easily to theoretical complications, and the supposed complexity forces oneself into strong philosophical hypotheses about the world and the language. Indeed, deflationists can even accuse pluralists of making the problem worse. By accepting multiple properties of truth, they run into many of the troubles that

traditional conceptions of truth have singularly taken, all at the same time.

It is with this situation in mind that deflationary views question the initial assumption: perhaps truth is not a complex property, maybe it has no nature or it is not a property at all. The classical philosophical proposals investigate truth as it was a property like "being magnetic" or "having causal power", namely a deep and complicated notion that deserves a deep and complicated account. Deflationism claims we do not need any of that. The mystery around truth is not to be dispelled because there is no mystery. We should "deflate" truth and look at it as a simple and trivial property lacking any deep metaphysical nature. The ambition of deflationism is to erase any complex view and speculative debates by erasing any complexity and metaphysical complications from the notion under scrutiny. Contrary to traditional views, deflationism avoids, or at least tries to avoid, commitments to big metaphysical pictures.

So, what is exactly deflationism? there is not a single, definitive answer. The term "deflationism", in fact, indicates more a family of different approaches sharing core claims about truth, rather than a well defined conception. To have a grasp of these claims, and of what the members of this family are like, it is helpful to follow the development of deflationism from the beginning to contemporary versions. This is what we do in the next sections.

EARLY DEFLATIONARY PROPOSALS: REDUNDANCY AND PERFORMATIVE THEORIES OF TRUTH

At least for what concerns analytic philosophy, the first stone of deflationism has been probably laid down by

Gottlob Frege[8]. Frege notes that the sentence "the thought that five is a prime number is true" says nothing more than the simpler sentence "five is a prime number". It seems that one does not add anything to a thought by ascribing the property of truth to it. From this, Frege concludes that the relation between a thought and its truth is not like the relation between a subject and a predicate. It is natural then to wonder whether we are not dealing with something that can not be called a quality in the ordinary sense. In these simple reflections we can find the seeds from which all subsequent development of deflationism grows. Anyway, Frege is not an advocate of an authentic deflationary approach to truth. Putting aside the subtleties of an exact interpretation of Frege's view, it is customary to say that for him truth is an abstract object, rather than a property. Truth is the object every true sentence names. At the same time, we cannot say much about the nature of this object because it is undefinable. The undefinability of truth makes him closer to a primitivist approach than to a deflationary one. In a primitivist approach truth is conceived as a substantial property, like in traditional views, although it is possible to define it in terms of other notions, because it is a primitive and fundamental property that cannot be analysed.

While in some of the later writings of Wittgenstein[9] we can find other remarks on the equivalence between the ascription of truth to a proposition and the proposition itself, it is with Ramsey[10] that such reflections are eventually used to motivate a first clear deflationary proposal. In a little more than a page Ramsey describes what will be called "the redundancy theory of truth". In his view the fact that

---

[8] Frege 1918.

[9] Wittgenstein writes, for example: "For what does a proposition's being true mean? 'p' is true = p. (That is the answer)". Wittgenstein 1956 Appendix III, par.6.

[10] Ramsey 1927.

to assert the truth of a proposition adds nothing to that proposition means that there is not a separate problem of truth but just a linguistic muddle. If the sentence "it is true that Caesar was murdered" means the same as "Caesar was murdered", then the truth ascription occurring in such a context is redundant. Indeed, we could say the very same thing without mentioning truth at all. If so, truth seems redundant, as it seems eliminable without any difference in the expressed content. From this we could infer that truth has no meaning at all.

However, Ramsey also notices that there are more problematic cases. Sometimes truth is ascribed to propositions that are not explicitly given, but only indirectly mentioned, like in "everything the Pope says is true". Henceforth such cases will be called *blind ascriptions* of truth. In such cases, we cannot just eliminate the ascription of truth, as we did above, since the result would be: "everything the Pope says", which clearly has a different meaning, and it is not even a complete sentence. However, what we mean can be rephrased as:

> "for every proposition p, if the Pope asserts p, then p is true".

Ramsey notes that here the predicate "is true" is just added to obey grammatical rules. We need it to get a well formed English sentence, but such an addition is not really necessary. Since "p" stands for a proposition, it already contains a verb. For the sake of simplicity Ramsey shows the point assuming that every proposition has only one relational form, like aRb. In this way the example could be rephrased thus:

> "for all aRb, if the Pope asserts aRb, then aRb".

In such a formulation the addition of "is true" is superfluous, so that truth can be eliminated again. However,

this analysis also shows that we can no longer conclude that a truth ascription is vacuous. Although when the proposition is explicitly given the truth predicate adds nothing, in blind ascriptions it has a meaning: it stands for an abbreviation of a complex quantified formula. Ramsey hence seems to have a sort of double reading[11] about the meaning of "is true" depending on what context it occurs in. In any case, in both contexts we can always avoid ascribing truth. If the notion of truth can always be eliminated, why, then, do we have such a predicate? According to Ramsey, having the word "true" allows rhetoric effects, like emphasis. The truth predicate enriches the language by enabling stylistic variety.

The idea of the rhetoric role of the truth predicate is stressed in another direction by the deflationary approach of Strawson[12]. If Ramsey holds that we do say something when we claim that a proposition p is true (although we say the same of p), Strawson argues that we do not *say* anything. Rather, we *do* something. When we ascribe truth we make a performative act, similar to other linguistic acts like promises or bets. The question then is: what do we do when we claim that something is true? According to Strawson, we confirm, we show agreement. Consider the example:

> Andrea says: "Vegetables do not taste good."
>
> Grazia replies: "It is true."

What Grazia does by saying "it is true" could have been done by saying "I agree" or by nodding. She does not describe some state of affairs; she acts.

Strawson is aware that we can also have different uses of the truth predicate, sometimes we express surprise or doubt, or we concede a point, like when we say "that cannot be true!" or "that's true but…". In order to vindicate all such

[11] See Kirkham 1992.
[12] Strawson 1949.

uses, then, Strawson must have not only one proposal but many. It must account for all performative uses of the truth predicate. This however is a problem for the entire approach. Indeed, it seems that what keeps together all these uses is some common meaning of "is true has" that goes beyond each single act. Moreover, it seems that such a common root cannot be in the performative level. the reason is that Strawson cannot explain arguments like:

if p is true, then q

p is true

then q.

In the premises "is true" seems to stand for different uses, serving different performative acts: hypothesizing versus stating. The argument then would equivocate and could not be valid. Moreover, if we just perform illocutive acts when we use the truth predicate, we do not really have an argument here, since there would be actions in place of authentic premises. Another, and probably the biggest problem for a performative approach is that, as Strawson himself concedes, some conditions in the world must hold for a correct use of "is true". We should not say that something is true unless it is the case. But this means that, after all, we do not use truth only to show agreement, but also to describe reality. Exactly what the performative view tries to avoid. A final problem for both Strawson and Ramsey is that they do not explain why we can apply the truth predicate only to declarative sentences. If the predicate is used to show agreement, why cannot we express our agreement with an order like "close the door!" by saying "it is true!"?

If Ramsey and Strawson provide early sketches of deflationary views of truth, the first detailed view, fully in the spirit of redundantism, is put forward by C.J.F.

Williams[13]. The root of his proposal is the construction Ramsey proposed for the interpretation of blind ascriptions in terms of quantified sentences. In the same spirit, Williams claims that he is able to give an analysis of every occurrence of the truth predicate just in terms of quantification, identity and conjunction. The truth predicate can thus always be explained away. Using such a reconstruction he concludes that "is true" is not a real predicate, and a fortiori it does not stand for a property. The first version of Williams' proposal is the same as Ramsey's. A sentence like "what the Pope said is true" is rephrased as "there is a p, such that the Pope stated that p and p". From this analysis Williams wants to deny that truth is a property by denying the existence of truth bearers. His argument focuses on the right side

there is a p such that _____ states that p and p.

Here Williams suggests that the expressions that should fill the gap must be incomplete symbols like "what the postman brought". These expressions are not names, in fact they can be negated ("what the postman did not bring" or "what the Pope did not say"), while names cannot. Since incomplete symbols do not name anything we have no entity with the property of being true. There are not truth bearers, therefore there is not a property of truth. This argument, however, is clearly unconvincing, since such expressions are complex descriptions that can denote objects even if they are not names. Moreover, in the case of explicit ascriptions of truth like ""snow is white" is true" we do have names for truth bearers, since we have quotation marks just to be able to form names for sentences.

Williams has also another argument against the idea that truth is a property. The expression "what the Pope said" involves the claim that the Pope said a single thing. Thus, we ought to reformulate the claim like:

[13] Williams C.J.F. 1976.

> A: there is a p, such that for every q, ((p is identical with q iff the Pope said that q) and p)

Compare now it with the sentence "what the Pope said is believed by Bob", that has a logical form like:

> B: there is a p, such that for every q, ((p is identical with q iff the Pope said that q) and Bob believes that p)

If we drop "Bob believes that" from B we obtain exactly the paraphrase proposed in A. Now, since in B there is no mention of truth we have no reason to claim that there is something corresponding to truth in A either. Also this argument, however, is unconvincing since it is questionable that we are allowed to do such manipulations and erase symbols at will[14].

In his final version, Williams' reading of a blind ascription has an even more complicated logical form. "What the Pope said is true" states two things:

1. the Pope said exactly one thing
2. for every r, if the Pope said that r, then r

So we get:

> F: there is a p, such that for every q, ((p is identical with q) iff the Pope said that q) and for every r (if the Pope said r then r).

Williams' final analysis is not only complicated, it is also a particular interpretation of a very particular sentence. It is not immediate how to extend such a proposal to a general view. For example, consider the sentence "Goldbach's conjecture is true" or ""snow is white" is true": it is inappropriate to use the verb "to say" both referred

[14] Apparently, in B, "p" occurs in term position whereas in P it occurs in sentence position, so that the argument seems to equivocate. To tame the problem, Williams' propose a non standard way to split that-clauses.

to a person and to a conjecture (or to a belief, a thought, a sentence, an hypothesis…). Conjectures do not speak. Apparently Williams is committed to a different analysis for every different truth ascription. But there is something even worse. If we try to get a general definition from F, we get something like:

> for every x [x is true = there is a p, such that for every q, ((p is identical with q) iff x says that q) and for every r (if x says r then r)].

Here the variable "x" in the left side must vary over a domain of some sort of entity, but this fact is incompatible with the claim that there are not such things as truth bearers. A final serious problem is whether and how the quantification used by Williams to bind variables really make sense. Although Williams claims that his proposal provides a redundancy view of truth, it does not seem so. The quantifiers Williams use apparently cover both singular terms occurrences (like in "p is identical to q") and propositional ones (like in "then r"). For the sake of charity, I pretended that this aspect could be easily ironed out, but it is actually problematic. "Is true" cannot be simply eliminated from English; it can be eliminated, at most, only from English *plus the machinery for propositional quantification.* The issue is indeed crucial for a deflationary treatment of the truth predicate and it is explicitly addressed, in a more systematic and convincing way, by the prosentential conceptions of truth.

THE PROSENTENTIAL CONCEPTION OF TRUTH

The idea of adding specific variables for sentential expressions, working for sentences like pronouns do for names, can be found in several authors such as Brentano, Lesniewski and Prior. Are we forced to add such variables or

are there *prosentences* in natural languages? Is propositional quantification already present in natural language? Dorothy Grover[15] thinks so.[16] Consider:

1. Sara likes apples, *she* likes pears too.

2. Paul brought me a present, *it* was a surprise.

3. Mary believes aliens exist, but I do not believe *it*.

In 1) "she" refers to Sara, in 2) "it" refers to Paul's having brought me a present and in 3) "it" refers to Mary's belief. These are called anaphoric uses of pronouns. In such examples we could avoid pronouns at all, for instance, 1) can be turned into:

1b) Sara likes apples, Sara likes pears too.

Beside anaphoric uses there are quantificational uses[17], like:

1) Every positive number is such that if *it* is even, you get an odd number if you add 1 to *it*.

In these contexts pronouns do not work as normal referring expressions. We cannot substitute them without a change in meaning:

4b) Every positive number is such that if every *positive number* is even, you get an odd number if you add 1 to every *positive number*.

Therefore, in such cases pronouns cannot be just avoided.

As pronouns for nouns, there are also proverbs for verbs:

---

[15] Grover 1992; Grover, D., J. Camp, and N. Belnap 1975

[16] Probably Brentano was the first to point out the existence of prosentences in natural languages considering the example of "yes". See Künne 2003.

[17] There are also other uses, like laziness uses, where pronouns can also be avoided.

> 1) If you decide to go, so *do* I.
>
> Analogously, there are also proadverbs and proadjectives.
>
> What Grover crucially claims is that there are also prosentences for sentences. Consider:
>
> 2) Snow is white. *That is true* but it rarely looks white in Pittsburgh.
>
> 7) Mary said that aliens exist, and I believe that *it is true*.

In these examples "that is true" and "it is true" work like anaphoric prosentences. We can rephrase 7) as

> 7b) Mary said that aliens exist, and I believe that aliens exist.

In these contexts prosentences inherit their references from other expressions, in the same way pronouns do with respect to names. In the same sense in which pronouns have not an independent meaning on their own, prosentences have not either.

At this point the basic idea of prosententialism can be introduced. Prosententialists claim that a certain fragment of English, in which "true" occurs only as a part of prosentences, has the same expressive power of entire English. This is very close to the spirit of a redundancy theory of truth, but prosententialists do not claim that "that is true" can be eliminated without any expressive loss. In fact, although "that is true" has the same content of the sentence it stands for, it differs under pragmatic considerations. Consider 7), by using "it is true" one does not only repeat what Mary said, one also recognizes that Mary has already said that. Another reason to avoid a pure redundancy view is that not every use of a prosentence is anaphoric. As in the

case of pronouns, there are also quantificational uses where prosentences cannot be avoided.

Beside the fact that "that is true" and "it is true" have no independent meaning, prosententialists claim that "is true" lacks content for another and more radical reason. They contend that when "that is true"and "it is true" occur as prosentences, the expression "that" or "it" only serve to form complex expressions and they have no meaning by themselves, not even as pronouns. In this context "that" or "it" is a mere syncategorematic term that occurs as a part of a whole complex expression that, in turn, can have a meaning. The same holds for "is true". Apparently "is true" is a predicate, but this is just an illusion rising from grammatical appearances. Also "is true" is a syncategorematic expression occurring in more complex and meaningful expressions. Accordingly, "is true" has no more content than the letter "d" in "dog". "Is true" true might be a predicate for superficial grammar, but semantically it is part of a semantic atom.[18]

Since, however, in English "is true" is not only used in "that is true" and "it is true", prosententialists must explain how to reduce such other constructions to purely prosentential ones. Let us start with simple cases of explicit ascriptions like ""snow is white" is true". Here the paraphrase is a straightforward application of a prosentence:

EA: Snow is white. *That is true.*

Generalized ascriptions are obviously harder, but they can be treated, as expected, in a similar way to Ramsey's solution. For example, "everything the Pope says is true" can be analysed as:

[18] It should be noted, however, that Grover often refers to "is true" as a predicate, and she is also open to the idea that truth is a property (Grover 1992, p. 22). Indeed, the semantics of "is true" is not completely clear. Grover, in any case, holds that even if an extension is sufficient for a property, it is not sufficient for an interesting property

> GA: Everything the Pope says, if the Pope says it, then *it is true*.

Where we have an example of a quantificational use of a prosentence.

Since in quantificational contexts prosentences work like variables for possible substitutions, Grover has reasons to motivate a logical analysis that replace "that is true" and "it is true" with propositional variables and substitutional quantifiers. Substitutional quantification[19] is a special kind of quantification that differs from standard objectual quantification. Consider "∃x(x is material)". According to the standard reading this sentence is true if (in the intended domain) there is an object that is in the extension of the predicate "is material". According to substitutional quantification, instead, variables do not vary over a domain of objects. They are associated to a set of suitable expressions that can be substituted to the expression "x" in the formula "x is material". In a substitutional reading of quantifiers "∃x(x is material)" is true if and only if there is at least one expression that can be substituted to "x" and such that the obtained sentence is (grammatical and) true. Deflationists meet substitutional quantification often along their way. We do not deepen the problem here, which is complex and fairly technical. What is worth noticing, however, is that it is debatable whether an account of substitutional quantification can be given without employing the notion of truth. Be it as it may, prosententialists do not aim at explaining truth away, but only at reducing the uses of a truth predicate to philosophically unproblematic and somehow shallow ones.

The kind of paraphrases sketched above is not only an interesting reduction of different uses of "is true" to purely prosentential uses. Grover, in fact, claims that there are no

[19] See, for instance, Dunn and Belnap 1968.

other uses in English apart from prosentential ones. "Is true" occurs in natural English only as a syncategorematic part of the prosentences "that is true" and "it is true". Different uses are only illusions due to surface grammar. Grover's analysis reveals the authentic deep logical form of such constructions. The moral of the story is not only that every use of the (apparent) truth predicate is equivalent to some prosentential use of it, but also that "true" is only and always a part of prosentences. The philosophical mystery of truth evaporates.

That "is true" is a not autonomous part of bigger prosentences is a thesis that leads prosententialists to troubles. How to analyse variants of the form "it will be true", "it could be true" and the like? Grover eventually admits that the deep structure of the language contains a big number of different prosentential operators. However, she has to deny that these operators have some shared grammatical structure, otherwise it seems that "is true" would have some kind of independent content, after all. Since such a conclusion is clearly not very convincing, other prostentialists, like Robert Brandom[20]. proposed different strategies. Brandom's version abandons the idea that "that is true" and "it is true" are not composed expressions, although , semantically, "is true" is not a real predicate. Truth ascriptions are not of subject/predicate forms. "Is true" is an expression that (semantically) works like a *prosentence forming operator*, and only superficially resembles a predicate. Namely, the truth predicate is an expression that can be combined with any kind of referring expression in order to get a prosentence. This move simplifies the analysis of blind ascriptions like "Goldbach's conjecture is true". In Grover's perspective it involves a quantified formula whereas in Brandom's it is immediately referring to the right proposition. Here "is

[20] Brandom 1994.

true" is combined with the definite description "Goldbach's conjecture" so that a prosentence standing for the conjecture of Goldbach is immediately obtained. Even the problem of prosentences like "it could be true" is easily solved. However the price for this is quite high. The moment Brandom allows that every referring expression can form prosentences when combined with a prosentence-forming-operator, he can no longer maintain that prosentences have no more content than the sentences they are about. Referring expressions in fact can involve richer contents. Consider: "what that bloody criminal, good for nothing man said is true[21]". Here, we do not express only a reference to a sentence asserted by someone; we apprehend also information about who asserts it. We know that he is a "bloody criminal, good for nothing man". Brandom might reply that this depends not on the truth predicate but on the expression it is combined with. However, it still seems that the entire manoeuvre misses its initial motivation. The mere prosentatialist analysis of truth seems less convincing. A second problem comes exactly from the admission that "is true" combines with referring expressions. Although this allows to fix the problems of standard Grover's prosententialism, it opens the doors to the idea that, after all, the truth predicate is a predicate both grammatically and semantically. If truth cannot be shown to be redundant or to be a bogus predicate, deflationists must deflate the property of truth in some other way. This is what, through Tarski, leads to contemporary deflationism.

FROM TARSKIAN SENTENCES TO CONTEMPO-

[21] A possible reply could be that this difference holds only at some pragmatic level.

RARY DEFLATIONARY CONCEPTIONS OF TRUTH

If previous proposals like Williams' redundancy theory and prosententialism have their roots in Ramsey's analysis of blind ascriptions, now we focus on explicit ascriptions. The basic idea is trying to explain everything about truth with some sort of equivalence between an explicit truth ascription and the proposition truth is ascribed to. The relevance of such equivalences arrives to modern deflationism through the classical work of the famous Polish logician Alfred Tarski[22]. Tarski searched for a definition of truth that could serve as a pivotal notion to clarify basic issues in mathematical logic. In order to do that he needed a definition satisfying two fundamental requirements. The first requirement is consistency. The possibility of a contradiction deriving from the liar paradox[23] is a big problem for any attempt at defining truth. To give a consistent account, however, it is clearly not enough to ensure that we have really defined truth rather than some notions in its vicinity. To avoid this a second requirement is needed. How could we be sure that our definition is an authentic definition of truth? The idea of Tarski is simple and brilliant. Everyone can have deeply different convictions about truth and its nature, but we all seem to agree on sentences like:

> "snow is white" is true (in English) if and only if snow is white
>
> "grass is green" is true (in English) if and only if grass is green
>
> "sky is purple" is true (in English) if and only if sky is purple
>
> …

[22] Tarski 1956.

[23] For more about the liar paradox and Tarski's treatment see the next chapter.

The idea can be generalized with the following schema:

TS: the sentence N is true in L if and only if S.

where "L" stands for the language we want to define truth for (called the *object language*), "N" stands for a name of a sentence in L and "S" stands for the translation of the sentence named by N in the language in which the schema and our theory (called the *meta language*) is formulated. When dealing with simple examples like the ones above, the addition of specifications about languages is not necessary, since the object language and the meta-language are the same: English. In general, however, this aspect is important. Such a distinction, in fact, is an essential part of the solution proposed by Tarski to solve the liar paradox and to avoid inconsistency. Also the notion of translation used in the schema is not strictly necessary if we deal with a single language, but it is important that the language in which the theory is formulated (the meta-language) contain a different name for every sentence of the object language. In natural language such names can be obtained simply by putting a sentence between quotation marks.

Apart from such technical notes, that we naturally accept those examples as a minimal condition on truth is a very likely hypothesis. But how can we turn this idea into the desired second requirement? Tarski proposed that if our definition of truth is adequate we should be able to prove[24], from that definition, all instances of the schema TS for the language we want to define truth for. If the notion defined respects this requirement it has the same extension of the intuitive notion of truth. Tarski turned this informal strategy into his famous Convention T[25], which is an explicit

[24] It is often said that we should prove all instances by purely logical means. This is not completely right in a Tarskian view, since Tarski also uses set theoretic resources. Probably it is more appropriate to say "by purely mathematical means".

[25] Note that "T" stands for "Truth" and not for "Tarski".

formulation of a perfectly acceptable adequacy condition.

It is easy to notice that these Tarskian biconditionals[26] share the same intuition behind Frege and Ramsey's initial speculations. When they notice the equivalence between an ascription of truth to a certain proposition and the proposition itself, they have in mind something similar to what Tarski spells out. Since, at first sight, such proposals seem to be the same, it is worth reflecting on the differences. First of all, both Frege and Ramsey take the truth bearers to be propositions or thoughts, while Tarski treats "is true" as a predicate for sentences. The reasons of Tarski's choice are probably two. On the one hand sentences allow us to avoid commitments to dubious entities like propositions and the subtle philosophical questions they raise. On the other hand, sentences have a clear syntactic structure on which Tarski can base the recursive machinery characterizing his own definition of truth, which should be noted, is not just given by TS. Another important difference concerns the type of equivalence at stake. The fathers of deflationism claim that "is true that snow is white" and "snow is white" have the same content or that one expression means no more than the other. For them, the equivalence is a very strong one. Indeed, the two constructions could be taken to be synonymous, adopting a schema like:

it is true that p $=_{syn}$ p

where "$=_{syn}$" indicates that the expressions at its flanks have the same meaning. It is from such a strong identification that the truth predicate can be argued to be redundant. Since the two expressions have the same meaning, we

[26] This kind of biconditionals for truth is known under different names: T-sentences, T-equivalences, Tarskian equivalences, Tarskian biconditionals, etc.. All of these often stand for different variants of the same basic idea. We avoid extreme rigour and leave the distinction to the reader.

can conclude that "is true that'' does not add anything. Tarski prefers a weaker option. For him, the equivalence holding between the two expressions is a merely material equivalence:

"p" is true ↔ p

Accordingly, a truth ascription to a sentence and the sentence itself have the same truth value. This idea, by itself, does not imply anything about the meaning of such expressions, so that he is not forced to adopt a redundantist position. ""p" is true" and "p" can have different meanings and the truth predicate could be not vacuous. The aim of Tarski indeed is to find a quite neutral criterion that would be acceptable to a wider range of philosophical views.

QUINE AND THE BIRTH OF CONTEMPORARY DEFLATIONISM

The first attempt at using Tarskian equivalences as the fundamental source to explain the nature of truth is proposed by Quine[27]. The key phenomenon revealed by such truth equivalences is, for Quine, that of *semantic ascent*. When we declare a sentence true, against the appearances, we do not speak really about a sentence but about the world, although in an indirect way. By calling the sentence "snow is white" true, we call snow white, says Quine[28]. The truth predicate is just a reminder that, although we speak of sentences, our eyes are on the world. The sentence ""snow is white" is true" is not about sentences, whereas "snow is white" is about snow: both are about snow. What tarskian biconditionals mean, according to Quine, is just this. If quoting a sentence gives us a name of that sentence, so that we can turn our

[27] Quine 1970, but see also Leeds 1978.
[28] Quine 1970.

discourse from world to language, the truth predicate erases the effect of quotation bringing the discourse back to the world. This is why Quine claims that "the truth predicate is a device for disquotation".

Why should we have such a device? In the answer that Quine gives to this question we find the mark characterising all subsequent forms of modern deflationism. Quine shows that by using tarskian equivalences and semantic ascent via a disquotation device, we can explain blind ascriptions too. Blind ascriptions are what forced Ramsey and others to complicate to introduce higher order quantifiers and complicate the early deflationist account of truth. All of this, Quine points out, can be avoided. Having a truth predicate which enables us to erase the effect of quotation, in fact, is extremely useful in contexts where technical or practical complications demand us to mention sentences, even if it is the world we want to talk about. This is the case, for instance, of certain generalizations. Suppose we were to assert all the sentences of the form "p or not p", namely the conjunction of all sentences obtained by substituting English sentences to "p", like:

C: (snow is white or snow is not white) and (grass is green or grass is not green) and (sky is blue or sky is not blue) and …

We clearly do not have direct means to assert such an infinitely long conjunction explicitly. In particular, we are not allowed to generalize using standard first order variables writing:

for every x, x or not x.

First order variables are objectual variables that must occupy name places, but here the variables occur in sentential positions. A solution, we know, could be the introduction of sentential quantifiers and variables, using substitutional quantifiers for instance. However once the disquotational

feature of the truth predicate is acknowledged, a simpler option becomes available. Thanks to the disquotational feature of the truth predicate we know that ""p" is true" is equivalent to p for any sentence p; thus, making all the relevant substitutions we can get an equivalent version of the infinite conjunction C, in the following way:

D: ("snow is white or snow is not white" is true) and ("grass is green or grass is not green" is true) and ("sky is blue or sky is not blue" is true) and …

At this point standard objectual quantification can be used:

for every x of the form (p or not p) x is true.

The disquotational nature of the truth predicate and its ability to allow standard objectual quantification over sentences is remarkable. It provides one of the main strengths of modern deflationist views.

CONTEMPORARY DEFLATIONISM: FIELD, HORWICH, SOAMES

An explicit formulation of Quine's ideas has been spelled out by Hartry Field who, together with Paul Horwich, is one of the chief advocates of contemporary deflationism. Field's version of deflationism[29] is focused on what he calls *pure disquotational truth*. The idea of the disquotational role of the truth predicate is essentially the same of Quine. Field however takes the predicate of pure disquotational truth (henceforth PD-truth) to be a predicate of utterances instead of sentences. In second place, in his proposal, the predicate can be applied only to utterances of which the speaker has some understanding. Every speaker, then, has her own PD-truth predicate for her own idiolect, which could differ, less

[29] Field has proposed different versions of deflationism. Here we refer to its original and typical version. Field 1986 and Field 1994a.

or more, from that of the other speakers. Everyone has a PD-truth predicate for how she understands it. The schema privileged by Field is:

for utterances u, the utterance that U is PD-true is cognitively equivalent to u.

The relation of equivalence in question is not material or analytic, but cognitive. According to Field two expressions are cognitively equivalent if the inferential procedures they allow are the same and they can be substituted one another without any difference in the inferential role.

This approach permits the application of Quine's analysis only into the idiolect of a speaker, which is a problem. For instance, suppose that Mary, who speaks only English (or better, her version of English), is a catholic that trusts the Pope completely, and says:

P: "Everything the Pope says is true".

Since the Pope can also speak German, which Mary does not understand, there are utterances of the Pope that Mary does not understand and to which she can not apply her PD-truth predicate. This is unfortunate, since Mary is sincerely convinced that everything the Pope says is true, even when he speaks German, so she would like to cover those utterances too. Field argues that it is possible to do justice to these cases in terms of PD-truth. When Mary states that P, she is conjecturing that for everything the Pope says there exists a good translation of what the Pope says in her own idiolect and it is this translation that is PD-true. In the final version then we have two notions of truth: a PD-truth, restricted to utterances of the idiolect of each speaker, and a notion of truth derived from this notion that can be applied also to utterances the speaker does not understand.

Such a complication is a consequence of the choice of utterances of a particular idiolect as truth bearers, which make the corresponding notion of truth restricted to a

singular language. Things would be different if we took the truth predicate to be a predicate of extralinguistic entities like propositions. This is the choice favoured by Horwich[30], the author of the most systematic and exhaustive account of a deflationary theory of truth. To be precise Horwich calls his own proposal "minimal theory of truth" in order to identify his version among the other deflationist theories. The minimal theory is very simple. It consists of all infinite axioms given by instantiation of the schema:

MT: the proposition that p is true if and only if p.

where p is whatever proposition expressible in a possible language. The schema holds also for those propositions that are not expressible in our language, English, or in any other current language. Since also for those inexpressible propositions there is an axiom in MT, some of the axioms of MT are not expressible. The instances of MT are not only material equivalences; they are also necessary and known a-priori. Horwich claims that:

1. MT can explain every fact involving truth (possibly joined with theories of different subjects)

2. The mastery of the concept of truth by a speaker consists in her disposition to accept any instance of MT.

Horwich's proposal is characterized by three main theses that make his position paradigmatic. First of all, he gives an account of the function of the truth predicate by developing the idea of Quine and applying it to a great number of cases. Horwich also stresses that the truth predicate enables us to avoid the syntactical and semantical complications of substitutional quantification. The second thesis concerns the account of the concept "true" and our use of the word. According to Horwich, MT gives the best explanation of our use of the truth predicate and (at least assuming a use theory of meaning) of the concept itself.

[30] Horwich 1998b.

This thesis is justified by the difficulty of finding contexts in which the notion of truth is involved but such that they cannot be explained in terms of the equivalences of MT. The last thesis is the flag of deflationism: it is the idea that truth lacks a substantial nature, so that it is helpless to try to define truth in terms of other notions. All explanations involving truth just need instances of MT.

Although the version based on propositions seems superior to the one based on sentences or utterances, the choice between such alternatives constitutes a dilemma:

1. if deflationism is construed in accordance to propositionalism it is trivial;
2. if deflationism is construed in accordance to sententialism it is false.

Consider the following case:

W: *snow is white* is true if and only if snow is white.

If we think that the left side of W is about the sentence "snow is white" then if W states a necessary claim, and as such W is false. The reason is that for the truth of the sentence "snow is white" it is not enough that snow is white. The sentence "snow is white" must also mean that snow is white and this is ignored by W. On the other hand, if on the left side there is a proposition, then the whole claim seems trivial, since as long as a proposition is individuated by its truth conditions, it can be defined to be *true* just in the case that snow is white. A way to circumvent this problem could be to consider not syntactical and meaningless sentences, but interpreted meaningful sentences. The version of the schema proposed by Scott Soames[31] goes in this direction:

S: if s means in L that P, then s is true if and only if P.

(where s is a sentence, L is a language and P is the proposition that s expresses according to L). Soames points

[31] Soames 1999.

out that every competent speaker would be ready to assert the corresponding instances of S, so it seems to be an essential part of our mastery of the concept of truth. In Soames' view, such instances are trivial, a priori and necessary.

S, however, makes explicit appeal to the notion of meaning to explain the notion of truth. This leads us to another problem afflicting deflationism and already noted by Michael Dummett[32]. If deflationists want to use the notion of meaning to explain truth, they seem to be forced to adopt a theory of meaning that is not based on truth. In particular, deflationists should abandon the traditional truth conditional semantics. However, rather than considering this a fatal objection, Horwich and Field consider it as a consequence of deflationism to take on board. Accordingly, these authors have proposed versions of inferential or use theory of meaning. It is worth noticing, however, that the problem seems to be more complicated[33]. On the one hand all conceptions of truth need the notion of meaning one way or another, so that the danger of circularity is shared also by rival views. On the other hand, to correctly evaluate the problem we should know what explicative role a deflationary theory cannot play. The notion of deflationary truth can do many jobs, as deflationists have shown, allowing, for instance, a certain kind of generalization. If truth only had deflationistically acceptable roles in a truth conditional semantics, then also a deflationist could adopt it.

THE CORE OF DEFLATIONISM

Although deflationism has been proposed in different versions, the analogies among such views are deep enough

[32] Dummett 1959.

[33] See Damnjanovic and Stoljar 2014, and Bar-On, D., Horisk, C. and Lycan, W.G. 2005.

to permit the individuation of a core of shared principles. When philosophers speak of deflationism, in fact, they often mean to refer to a generic conception of truth committed to a limited set of claims, without having in mind a particular version. By conforming to this widespread practice, we do not mean to minimize or to neglect the differences of each single theory. We do that just for the sake of simplicity and to allow a general treatment.

There are three main principles characterizing contemporary deflationary views:

1. T-sentences (in some form) govern the truth predicate and explain every fact involving truth;

2. The notion of truth exists only to serve certain logico-grammatical purposes. Namely, it is a (disquotational) device to form a certain kind of generalizations expressing infinite conjunctions and disjunctions;

3. Truth is not a substantial property.

The point 1., even in its different formulations, is the most evident feature of a deflationist approach. The centrality of biconditionals makes a deflationary view a conception based on very simple if not trivial principles. Notice, however, that point 1. is not enough to characterize a deflationist position alone: there exist, in fact, other theories that recognize the absolute importance of tarskian biconditionals but such that their advocates reject deflationism, for example the Revision Theory proposed by Gupta and Belnap[34].

A more characteristic point is the second according to which the biconditionals are used to explain the function of the truth predicate to express infinite conjunctions and disjunctions. Thanks to this function we can express generalizations involving a great number of sentences, like when we say "all the axioms of Peano Arithmetics are true". As we have seen, the idea of this mechanism goes back

[34] Gupta and Belanp 1993.

to a proposal of Quine and it is grounded on T-sentences, so that this second claim is grounded on the previous one. Also in this case, though, the claim is not enough to characterize deflationism, since other theories can vindicate the disquotational role of the truth predicate and its logical function. Whatever view includes the biconditionals could do the same.

The third thesis states that truth is a very special property, since it is insubstantial and it is not definable. This thesis is the crucial mark of deflationism. Unfortunately, this is the hardest deflationist idea to make sense of. The idea that truth is not a definable property seems to drag deflationism toward a primitivism. The whole set of T-sentences in fact does not provide an explicit definition in the standard sense. Rather T-sentences give something like an implicit definition[35] treating truth as a primitive notion and showing what principles govern it. Why, then, is not deflationism a simple primitivist theory? The reason lies in the thesis of the insubstantiality of truth. The other two ideas - the centrality of T-sentences and the logical function of the truth predicate - can easily find room also in a primitivist conception of truth. Thus, without the third claim, deflationism could also give rise to a primitivist proposal. Of course, there is a sense in which the intuition behind a primitivist view is very different from the one sustaining deflationism. After all, what could be more inflationary than thinking that truth is a property that is so fundamental to be unanalysable? Moore's primitivist view about good has been viewed in this light. However, the fact that one view has a different inspiration from another does not mean that, at bottom, they cannot be equivalent views. In other words, the idea that deflationary truth plays an important logical role does not distinguish

[35] Here we speak of "implicit definition" in the sense in which axioms implicitly define a notion in an Hilbertian sense. This, however, can be a subtle and complicated point, see Bays 2006 and Halbach 1999b.

the metaphysics of deflationism from the metaphysics of primitivism; and it is the metaphysics that is at stake here.[36] To clearly distinguish deflationism from primitivism, the crucial point is exactly the third deflationary thesis: truth is not a substantial property. But what does this mean? If it is not enough to point out that the notion of truth is not definable and that it is not analysable, how can we clarify the insubstantiality of truth?

This anti-substantialist stance is central to deflationism since its own birth, although its exact meaning has become harder and harder to understand. According to redundantism the truth predicate is considered vacuous and eliminable without expressive loss. If so it would be quite simple to argue that truth is not a substantial property, since the property of truth does not exist, and nothing is less substantial than something that does not exist. The performative proposal of Strawson can give the same direct explanation, since, again, the expression "is true" is not used to ascribe a property. The story becomes more complicated with prosententialism, which claims that truth cannot be eliminated without expressive loss. The expression for truth has a specific role: it occurs as a syncategorematic part of prosentences. However, Grover and other prosententialists can adopt a double move: on the one hand they can insist that "is true" is a bogus predicate. On the other hand, they can insist on the fact that a prosentence has no independent meaning, exactly as pronouns do not.

In all such cases deflationists have been able to claim that truth lacks a substantial nature by insisting that "is true" is not a real predicate, at least under a semantic point of view. Modern deflationists, however, have more troubles. They accept the centrality of T-sentences and they are naturally led to the thesis that the truth predicate

[36] Damnjanovic and Stoljar. 2014.

is an authentic semantic predicate. They cannot adopt the redundancy move either, because such a predicate is not always considered avoidable. According to modern deflationism truth is a predicate with a standard extension in a precise sense: T-sentences give us for every sentence a clause to decide whether the sentence in question is or is not into the extension of "is true". It is at this point that keeping insisting on the peculiar nature of truth becomes extremely difficult. On this difficult theme, for a long period of time deflationists have mostly proposed mere suggestions or slogans like "truth is a logical property". Indeed, the point is so obscure that a champion of deflationism like Field has admitted not to be clear enough as to what the insubstantiality of deflationary truth is supposed to mean.

Given the importance of the third claim to obtain a peculiar and clear characterization of deflationism, a precise account of the alleged insubstantiality of truth is indispensable. Only insubstantiality can vindicate the originality of the deflationist approach to truth and give means to resist the idea that the deflationism truth has just been progressively re-inflated. In this book we discuss at length one precise strategy to obtain such a clarification of the insubstantiality of deflationary truth, namely conservativeness.

CHAPTER TWO
FORMAL THEORIES OF TRUTH[37]

Not only is truth a notion of great philosophical interest, it is also a subject of scrutiny for logic and mathematics, where it is analysed by means of very complex formal tools. Historically, such an approach finds its official birth with the work of Alfred Tarski, who succeeded in showing how it is possible to define a truth predicate in a rigorous way, at least for formal languages with particular features. Doing so he swept away the suspicion that the notion was irreparably wasted by semantic paradoxes and irremediably incoherent. The paradox of the liar, the main semantic paradox, follows easily from principles that look absolutely innocent, and it seems to be strictly tied to the nature of truth. Consider the sentence:

L: the sentence L is false.

Such a sentence looks grammatically well-formed; it does not seem that there are ambiguities or category mistakes, and it looks perfectly meaningful. The sentence L seems to say of itself that it is false and nothing else. At first sight, the fact that it speaks about itself might appear worrying, but there are a lot of cases of self reference that

[37] Also when not explicitly reported, for what concerns axiomatic theories of truth, in the entire book, general reference is to Halbach 2011, and Cieslinski 2017.

are completely safe[38], for example:

I: the sentence I is an English sentence.

However, moving from L we can argue:

1. the sentence L is false (hypothesis)
2. the sentence L is true or false (bivalence)
3. suppose L is true
4. if L is true, then things are the way it says, so L is false (basic principle of truth)
5. then if L is true, L is true and false (by 3. and 4.)
6. suppose L is false
7. if L is false, then things are not the way it says, so L is true (principle of truth)
8. then if L is false, then L is true and false (by 6. and 7.)
9. L is true and false. (by or-elimination in 2.)

This argument, even in this simplified form, shows how easy it is to get an absurd conclusion (that a sentence is both true and false) just using apparently innocuous or well established principles. Hypothesis 1. is a matter of fact, 3. and 4., which I have called "principles of truth", represent the heart of our intuitions about the notion of truth and they could be reformulated in different ways. The other steps apply basic rules and laws of classical logic. This argument reveals that a serious problem lies hidden in some of the basic principles of our conceptual system. Taking logical rules and our simplest intuitions about the notion of truth together we are quickly led to an absurd conclusion.

If before Tarski the common attitude towards these problems was pessimistic and the notion of truth appeared to be irreparably incoherent, the Polish logician showed that it is possible to handle truth safely. A paradox free definition of truth was shown to be possible,

[38] It is often pointed out that the phenomenon of self-reference can be rebuilt also in formal frameworks in a rigorous way so that the legitimacy of it should be taken for granted. See Visser 1989.

at least in certain cases. In particular, Tarski focuses on artificial languages that meet the requirements of not being *semantically closed*. Where a language is not semantically closed if it cannot express its own semantics. More precisely, a non semantically closed language (namely, an open one) does not contain its own truth predicate. Note that in a language that is not semantically closed, it is impossible to construct a sentence like L, so that the liar paradox cannot even arise.

The notion of a semantically closed language is often confused with the distinction between an *object-language* (the language for which we want to define truth) and a *metalanguage* (the language in which we define truth). Often it is this distinction that is thought to be the tarskian solution to paradoxes. This is not correct: such a distinction is important only when we are interested in giving a definition of truth; it does not serve any purpose when we are only working to solve the paradox. Of course, the solution is an essential premise in order to permit a coherent definition of truth: if the languages were not semantically open, we would be victim of the liar again. The two issues, however, are different and it is worth keeping them distinct. The request that the language under scrutiny must be semantically open is what allows us to handle safe languages. The distinction between an object-language and a metalanguage becomes important only when we also want to give a definition of truth. This leads us to a further requirement that we must impose to be able to obtain a definition of truth: if the object language is really open, then it must not be possible to translate the metalanguage in it. Otherwise, the truth definition could be translated back into the object-language, making it closed after all.

Clearly, a coherent definition is not enough to guarantee that what we have found is a definition of the very intended notion we intended to characterize, namely truth. At this

point, however, tarskian biconditionals help us: if the definition we find enables us to deduce all the tarskian biconditionals for the object-language, then our definition is materially adequate and it captures a notion of truth[39].

The same kind of considerations clearly holds if we would like to give a definition of truth also for the metalanguage. In this case, we would need a meta-metalanguage, distinct from the previous metalanguage, in which we can build the new definition. In this way we obtain another truth predicate with a different extension (since it would not apply to sentences of the first object language but also to sentences of the metalanguage). The process can be iterated to obtain a definition of truth for the meta-metalanguage in a meta-meta-metalanguage and so on, yielding a hierarchy of truth predicates for languages of higher and higher orders.

The solution proposed by Tarski, although adequate for his goals, is, under certain respects, not fully satisfactory. In particular, it does not seem possible to extend the Tarskian approach to natural languages. On the one hand, natural languages seem to be semantically closed, so that they seem irreparably victim of the paradoxes. On the other hand, the existence of a hierarchy of different truth predicates does not seem confirmed in usual linguistic practice. In any case, thanks to Tarski, the notion of truth and the paradoxes have been shown to be liable to rigorous treatments, opening the way for further research.

It is with the work of another great logician, Saul Kripke, in the seventies, that became clear that some defects of the Tarskian proposal could be overcome. After Kripke, the work on the liar paradoxes rapidly increased, with refined technical results often little-known to non specialists. Kripke's project is based on two main ideas: a construction of an interpretation of a language containing its own truth

[39] See infra Chapter One.

predicate (thus semantically closed), and a solution of the paradox based on the abandonment of bivalence. The law of bivalence, the principle according to which every sentence is either true or false has a key role in the derivation of the liar paradox. By rejecting bivalence, and offering a suitable construction, Kripke shows how paradoxical sentences can be considered as neither true nor false.

WHAT IS TRUTH FOR?

Although work on semantic paradoxes treats truth as a subject of investigation, Tarski's main aim was to use truth to account for other notions. A solution to paradoxes was only a preliminary stage to obtain a notion suitable for other uses. The chief and now classical applications are the formulation of model theory in its modern form and the definition of logical consequence.

It is common in logic to consider a language in purely syntactical terms, as a set of sequences of certain symbols inductively constructed according to precise rules. We have, then, a set of symbols divided in categories (logical symbols, and, possibly, individual constants, constants for n-places predicates, individual variables…) and a set of rules telling us how to put these symbols together to obtain well formed formulas and sentences. Since the characterization is syntactic, the expressions of the language are not given a meaning yet. This makes the use of the word "language" a little odd compared with the common usage. If we want the expressions of the language in question to have a meaning we need to give them an interpretation. This is what is done by construing a *model* for the language. Usually a model M is identified with (at least) an ordered pair:

$$M = \langle D,I \rangle$$

where D is a domain of individuals and I is a function of interpretation that gives a meaning to the expressions of our language with regard to D. By specifying the model we determine what objects the individual constants stand for, what relations the predicative constants stand for, and so on. We give the language a semantics. It is clear that the same language might be liable to different interpretations. The same sequence of symbols can have different models, and thus different meanings. The same sentence, then, will be true under certain interpretations and false under others, splitting the possible models in two classes: the class of the models that make the sentence true, and the class of those that make it false. The process can be read also the other way round: given a certain model, we can divide the sentences of the language into two groups: the group of the sentences true in the model and the group of the sentences false in it. This very informal and quick presentation of the basic ideas of model theory should suffice to see the importance of having a rigorous definition of what truth-in-a-model is. Until Tarski, however, to call a sentence "true under an interpretation" was just a way to speak. Relying on his rigorous definition of truth, Tarski clarified the central notions involved and permitted an unprecedented growth and a solid foundation of model theory. The role of Tarskian definition of truth in this part of mathematics is now so deeply rooted that it is quite surprising that some important results in the field (like Lowenheim-Skolem's theorems) were proved when a rigorous definition of truth was not available yet.

The notion of truth, as characterized by Tarski, has an essential role also in the specification of logical consequence. Indeed it is only with Tarski that an adequate characterization of this notion and of logical truth has been possible. The definition, nowadays standard and described in every basic textbook of logic, employs the notion of

truth-in-a-model sketched above. A sentence φ is said to be a logical consequence of a set of sentences Γ (in symbols $\Gamma \vdash \varphi$) if and only if every model that makes the set of premises Γ true makes also the conclusion φ true. Where, again, the notion of a model making sentences true is defined in a rigorous way with the tarskian definition of truth.

The notion of truth, beside being used as a key concept to define other model theoretic notions, has been used and investigated also as a tool for analysis in proof theory. Here the truth predicate, governed by a set of axioms, is used to obtain interesting intertheoretic reductions[40]. The most common kind of theories that are liable to such reductions is provided by subsystems of second order arithmetic (analysis) on the one hand, and by axiomatic theories of truth on the other hand. The chance of reducing such subsystems to certain theories of truth allows, for instance, to translate sentences about numbers and sets of numbers into sentences speaking only of numbers. In this way, formulations involving ontological assumptions on sets can be translated into formulations that only presuppose semantical assumptions.

Apart from ontological motivations, such intertheoretic reductions are motivated also on the ground of (meta) mathematical reasons. One of these reasons can be, for instance, the goal of proving the consistency of a certain theory. If a theory can be reduced to another theory, which is already known to be consistent, then the former theory must be consistent too. Another, and related interest towards truth in proof theory is motivated by considerations about the foundation of mathematics. In particular, a field of research that is growing in recent years, that of *Reverse Mathematics*, aims at stating what kind of mathematical reasoning can be achieved in determined subsystems of analysis: what

[40] See Halbach 2000.

specific resources are needed to prove particular theorems. Since some of those subsystems are equivalent to certain theories of truth, larger and larger fragments of ordinary mathematics can be rebuilt in truth theories. If this is joined with the innocent ontological presuppositions a theory of truth seems to have, the attraction of such reductions becomes clear.

AXIOMATIC THEORIES OF TRUTH

So far we have been speaking of formal approaches to truth in a quite relaxed way, referring to any investigation using logico-mathematical tools. Such approaches can follow different ways and they can be elaborated in different formal frameworks. It is possible to give a definition of truth, as Tarski did, or to build interpretations of the truth predicate, yielding models. This probably is the most notorious approach, and it can be found exemplified in typical ways in the work of Kripke or in the Revision Theory of truth.[41] In proof theory we find another way to investigate truth: an approach based on axioms. According to this approach the truth predicate is treated as a new symbol of a formal language and axioms governing the behaviour of such a predicate are presented. Often such axiomatic theories are inspired by proposals developed in model-theoretic frameworks. Such model-theoretic characterizations often leave room for manoeuvre, so that different axiomatizations of the same model-theoretic proposal can be obtained. This is another case where intertheoretic reductions find application. If an axiomatization is able to define the specific resources used to define truth in a model-theoretic approach, this is evidence that the axiomatization of such

[41] Gupta and Belnap 1993.

interpretation is adequate[42].

There is a great number of axiomatic theories that have been proposed and investigated. They are often known by abbreviations: we have, for instance, T(PA) (the theory inspired by the Tarskian definition), KF (the theory inspired by Kripke's construction and proposed by Solomon Feferman[43]), or VF (a version of the Kripkian proposal developed by Andrea Cantini[44] using VanFraassen supervaluations), just to cite some.

Here we work, essentially, within an axiomatic approach, operating with (simple) axiomatic theories and using some of the formal tools that have been elaborated in this field of research. Moreover, we will consider only those theories, and the aspects of those theories, that are relevant for the debate on deflationism.

THE BASE THEORY

According to an axiomatic approach, truth is conceived as a predicate of certain objects, which in philosophy are usually called *truth bearers*. If we are to model axioms and rules governing the behaviour of the truth predicate and to investigate the differences between different axiomatizations, having an independent theory that describes the properties of the objects to which we want to ascribe truth is necessary. In philosophy the debate over the nature of truth bearers is currently open, and discussions about the identification of the correct truth bearers are often tied with philosophical issues. In logic it is a common practice to take truth bearers to be sentence-

[42] This is the case, for instance, of the equivalence between T(PA), the theory of truth inspired to the Tarskian definition, and ACA, a fragment of second order arithmetic (See Chapter Four).

[43] See for instance Feferman 1991.

[44] Cantini 1990.

types. The reasons are similar to the ones already mentioned about Tarski. Apart from general motivations (like the less controversial nature of sentences-types with respect to others and the less demanding metaphysical commitments sentences compared to proposition) there are motivations of formal convenience: sentence types of artificial languages have a grammatical structure that can be specified by precise inductive rules. In any case, the treatment given in an axiomatic theory is based on such a few assumptions that the question about the exact nature of truth bearers cannot only be put aside, but left open to some extent too. Whether truth bearers are propositions, sentences or utterances will make a little difference, as long as they satisfy some simple constraints.

The theory describing the features of the objects to which truth is ascribed isa syntax theory. Many different theories can be chosen as a syntax theory. Possible options are, for instance, Peano Arithmetics, a theory of concatenation, or set theory. The choice in favour of a syntax theory is often motivated with considerations beyond the simple desire of a syntax theory for a theory of truth[45]. Here we will sketch and use Peano Arithmetic (henceforth PA) as syntax theory for two reasons: PA is one of the more commonly used base theory and it is indispensable for our discussion.

PEANO ARITHMETIC

The language L_{PA}, in which PA is formulated, includes the usual symbols of first order logic with identity $\{\wedge,\vee,\neg,\rightarrow,\leftrightarrow,=,\forall,\exists,(,)\}$ and an infinite numerable set of individual variables $\{v_i \mid i \in \omega\}$. The formation rules for the set of well formed formulas are the usual ones and so are

[45] For instance, we could prefer the base theory to be finitary, choosing, for example, PRA (Primitive Recursive Arithmetic).

the conventions for parentheses; we will also write x, y, z... instead of $v_1, v_2, v_3 \ldots$ To these symbols we add the individual constant "0", a symbol for a one-place function "S", and two symbols for two-places functions "+" and "•". The formation rules are enriched in the following usual way:

i. if t is a term of L_{PA}, St is a term of L_{PA};

ii. if t and q are terms of L_{PA}, +tq is a term of L_{PA};

iii. if t and q are terms of L_{PA}, •tq is a term of L_{PA};

We adopt the convention according to which the symbols of functions "+" and "•" are interposed between the terms, so we write "t+q" and "t • q" instead of "+tq" and "•tq". The axioms and rules of PA include those of (classical) first order logic in one of its formulations (assuming a system of natural deduction in a Lemmon-style as general reference), and axioms for the new symbols. These specific axioms are:

P1. $\forall x \neg(Sx=0)$

P2. $\forall x \forall y(Sx=Sy \longrightarrow x=y)$

P3. $\forall x(x + 0 = x)$

P4. $\forall x \forall y(x + Sy = S(x + y)$

P5. $\forall x(x \bullet 0 = 0)$

P6. $\forall x \forall y(x \bullet Sy = x \bullet y + x)$

plus the axiom schema:

P7. $F(0) \wedge \forall x\,(F(x) \longrightarrow F(Sx)) \longrightarrow \forall x\, F(x)$

where F(x) is any formula in L_{PA} with exactly x free.

Notice that the axioms of PA are infinite. Since P7. is a schema - called the induction schema[46]-, it has infinite instances.

These axioms are the first order formalization of the axioms originally proposed by Peano/Dedekind to obtain an axiomatization of arithmetic. P1. states that zero is not a successor of any number, P2. states that if two numbers have the same successor then they are equal. P3., P4., P5 and P6.

[46] It can be considered as a rule or as a list. This can be an important difference when the language is extended with new symbols.

characterize the operations of addition and multiplication.

The schema of induction deserves some more words. The schema proposed here is the first order version of the axiom that Peano originally stated as: "every property such that zero has that property and such that if a number has the property then its successor has the property too, is a property of every natural number." This means that if a set of natural numbers includes zero and it is closed under the successor operation, then it includes every natural number. Formulated in this way the axiom speaks of properties of natural numbers and it quantifies over such properties. A first order language has no expressive resources to state the schema, since it does not include variables for properties and quantifiers able to bind them. All we can do is mimicking this principle with the schema of induction, taking it to hold for every formula in the language of PA that defines a set of natural numbers. Accordingly, we have to imagine that there are infinite instances of the schema, one for every property expressible in L_{PA}.

The induction principle is applied in a systematic manner in normal mathematical and meta-mathematical practice, and it is essential for the proof of a great number of theorems. This principle states something deeply bound to the nature of natural numbers. Its importance is such that if anybody did not recognize the validity of the schema, we could arguably conclude that this person has not understood the nature of natural numbers. For such a reason the principle of induction seems to hold for every formula that identifies some subset of natural numbers.

THE ARITHMETIZATION OF SYNTAX[47]

As remarked, a syntax theory giving information about the properties of the objects to which we want to ascribe truth is deeply valuable. The fact that PA has been chosen as a privileged syntax theory, however, might look curious at first sight. PA has been explicitly formulated to grasp, in a first order theory, arithmetic, namely the theory of natural numbers. Apparently, it does not deal with sentences, propositions or the like. Why did we take PA then? The point is that an arithmetical theory, like PA, can be also seen as a theory of linguistic expressions. How this is possible has been shown by Kurt Gödel, who used and adjusted what is today known as "arithmetization of syntax". The arithmetization of syntax is a set of operations that enables us to treat natural numbers as codes of expressions, and syntactic properties of expressions as arithmetical sets. Although the technical machinery to this aim is quite complicated, the basic idea is, on the contrary, very simple. First, different natural numbers are made to correspond to different primitive symbols of the language L (connectives, constants, variables…). Second, a different unique natural number is assigned to the syntactic complex constructions of L, depending on the symbols (and their codes) occurring in such complex expressions. In other words, a numerical code is assigned to symbols and expressions of the language, and this process is arranged in such a way that to every symbol or string of symbols is assigned a different natural

[47] For what concerns the notation, I sacrifice extreme rigour in favour of perspicuity. For example, I often write quantifiers and connectives within the scope of the "T" predicate, although, to be precise, I should speak of functions from Gödel numbers to Gödel numbers representing the effect of applying those quantifiers and connectives. See Halbach 2011 for details.

number. Moreover, it is necessary that our attribution has an algorithmic nature, so that given a number we can find out the expression it codes, and given an expression we can recover the number it is coded from. Since a different natural number corresponds to each sentence, such a number can be considered as a name for that sentence. At this point a discourse about a set of expressions can be seen as a discourse about sets of numbers: the sets of codes (even called *Gödel numbers* or *Gödelians*) of such expressions.

Also L_{PA}, the language of Peano Arithmetic, can be coded in that way by attributing in a proper manner numbers to its symbols and sequences of symbols. Thus, the fact that PA has been built just to talk about natural numbers leads to the surprising consequence that PA is a theory that can, thanks to the arithmetization of syntax, speak also "about itself". As it is well-known, this is what allowed Kurt Gödel to construct his great proof of the incompleteness of PA[48].

If PA can indirectly speak about (its own) expressions by indirectly speaking about numbers, we have to clarify what syntactic properties and relations can be expressed in PA. A property of expressions can be seen in the present context as a simple set[49] of numbers: the set of the codes of expressions that have that property. The problem is then to define what numerical sets can be expressed inside PA. For this purpose we need to give a precise sense to the ability of PA to express relations. This clarification is based on the notion of *representability*. A n-ary relation R among natural numbers is representable in PA[50] if and only if there is a formula $\alpha(v_1,\ldots,v_n)$ in L_{PA} with exactly n free-variables, such

[48] To be more precise, Gödel did not work with PA but with a formal version of arithmetic inspired by Russell's *Principia Mathematica*.

[49] When we have n-placed relations we have sets of ordered n-ples instead.

[50] We restrict our attention to PA, but the definition can be extended to any formalization of arithmetic.

that, for every n-pla $^{<}n_1,...,n_n^{>}$, $^{<}n_1,...,n_n^{>}$ belongs to R if and only if[51] $\alpha(v_1/\ \mathbf{n_1}\ ,...,v_n/\mathbf{n_n})$ is a theorem of PA. Where "n_i" is a variable for natural numbers and the corresponding "$\mathbf{n_i}$" in bold type stands for the numeral[52] of "n_i". Such a definition can be easily extended to functions.

It can be shown that some simple arithmetical relations and functions (as numerical identity, multiplication…) can be represented inside PA. This is not very surprising since PA has been constructed to talk about numbers and the main operations on them. The interesting question is if and what other relations and functions are representable in PA. We can give a very precise answer to this question: every *recursive* relation (and function) is representable in PA. Thanks to the notion of *recursion* Gödel is able to specify the intuitive idea of decidability or computation. In general, by "computable function" we mean a function for which we have a procedure or a set of procedures that allows, in a finite number of elementary steps, to solve the function for any given argument. A function is decidable if there is an algorithm that always enables us to get the value of the function. Such pretheorical conception has been investigated under different approaches using the notion of Turing machine, proposed by Alan Turing, or the lambda calculus of Alonzo Church, to cite some notable examples. Since these proposals are equivalent, there is solid room to conclude that any of them is an adequate explanation of the intuitive concept of computation. The theory of recursive functions is one way to explain the notion.

Briefly[53], a function is recursive if it is constructed

---

[51] If we have only the left-to-right direction, we say that the relation is semi-representable.

[52] Or better, for the singular term that names n_i in L_{PA}.

[53] For sake of simplicity, we do not distinguish between primitive recursive functions and recursive functions (which include the operation of minimalization).

from certain initial functions by the application, for a finite number of times, of some base operations. The set of recursive functions is then a set defined by induction on the set of base functions and closed under certain specified operations. The idea is that the base functions are certainly computable, thus applying to them certain clearly computable operations we obtain functions that are still computable. The definition of recursive functions can be easily extended to sets: we will speak of recursive sets if the corresponding characteristic function (which is the function that for any argument tells whether the argument belongs to the set or not) is a recursive function. Similarly we speak of recursive properties if the set of objects that have that property is recursive.

We do not give details of the theory of recursion here, specifying what these initial functions and operations are. However, it is worth noticing that among recursive properties (properties that are representable in PA) there are many important syntactic properties. Indeed, it is possible to define in PA the sets of (codes of) expressions that have properties corresponding to the main syntactic categories (like being a sentence, or being a singular term). This fact follows easily from the procedure of coding. Also more complex properties and relations among expressions can be represented in PA. A very important property is that of *provability in PA*. Let Σ be a sequence of formulas in L_{PA} and φ a formula in L_{PA}; the relation according to which Σ is a proof in PA of φ is a recursive property. This means that if m is the code of the sequence Σ and n the code of φ, then a formula (be $Prov_{PA}(x,y)$ such a formula) that represents the relation of provability in PA can be defined in PA. In other words, $Prov_{PA}(m,n)$ is a theorem of PA if and only if m is (the code of) a proof in PA of the formula (coded by) n. Thus it can always be verified in an algorithmic way if a given sequence of formulas is a proof of a certain given formula. This does

not mean, though, that we can algorithmically build a proof for an arbitrary formula: the property of a formula of being provable in PA, of being a theorem of PA, is not a recursive property. The set of formulas for which there exists a proof in PA, which we can define as the property expressed by $\exists x \mathrm{Prov}_{\mathrm{PA}}(x,y)$ with y free, does not correspond to a decidable set, but only to a semi-decidable set. Such a set is recursively enumerable, since we have an effective procedure to list its members, but we do not have a decision procedure allowing us to determine for every formula whether it is a theorem of PA or not.

THE UNDEFINABILITY OF TRUTH

Thanks to the arithmetization of syntax, we can speak of expressions of a language and define some properties of those expressions using an arithmetical theory. It is natural then to wonder whether we are able to represent in PA also the property of being true. Namely, whether there exists a formula $\tau(x)$ in L_{PA} such that for every sentence in L_{PA}, PA proves $\tau(\ulcorner\varphi\urcorner)$ if and only if φ is true. Where "$\ulcorner\varphi\urcorner$" stands for the Gödel number of the sentence φ.

To be precise, so far we have treated L_{PA} from a purely syntactic point of view, so that speaking of the truth of a sentence of L_{PA} does not really make sense unless some interpretation giving our language a meaning is provided. In other words, a model must be specified. Since PA is a theory built to formalize arithmetic, it is natural to consider the truth of PA with regard to what is called the standard model of arithmetic $\mathbb{N}$. So reformulated, the question is whether there exists a formula $\tau(x)$ in L_{PA} such that for every sentence φ in L_{PA}

$\mathrm{PA} \vdash \tau(\ulcorner\varphi\urcorner)$ if and only if $\mathbb{N} \vDash \varphi$.

Unfortunately, as proved by Alfred Tarski, this is not

possible. Tarski, actually, proved an even more general result:

2.1 Tarski's theorem:

The set of Gödel numbers of arithmetical sentences that are true in the standard interpretation is not arithmetically definable.

A formula $\tau(x)$ in the language of arithmetic such that for every arithmetical sentence ψ,

$$\mathbb{N} \vDash \tau(\ulcorner\psi\urcorner) \leftrightarrow \psi$$

does not exist.

This theorem is not strictly about sentences in L_{PA} or representability in PA, but it is, generally, about arithmetical sentences and arithmetical definability[54]. What we refer to with these expressions is the arithmetical theory as it is commonly intended by mathematicians and common sense. Intuitive arithmetic is a not formalized theory and PA is a proper fragment of it. It follows that the set of arithmetical sentences that are true in the standard model is not recursive, a fact that would imply representability in PA. In fact, suppose it was recursive, then, we know that it would be representable in PA. However we can take PA to be a fragment of intuitive arithmetic, so it would be definable in intuitive arithmetic too. We know from Tarski's theorem that this is not the case, thus, it is neither representable in PA nor recursive. In other words, we know that "if truth is representable in PA, it is arithmetically definable", but we know, by Tarski's theorem, that it is not arithmetically definable so it is not representable in PA either.

Although it is not possible to represent in PA the truth

[54] See Boolos, Burgess and Jeffrey 2007.

of sentences in L_{PA}, nonetheless it is possible to go very close to this, as the next theorem shows.

2.2 Theorem[55]:

The set Tn of (codes of) sentences in L_{PA}, with logical complexity less or equal to n, true in the standard model is arithmetically definable. (Where the logical complexity is the number of occurrences of logical operators).

Note, in particular, that the set T_0 of the atomic sentences in L_{PA} that are true in the standard model $\mathbb{N}$ is indeed recursive.

Hence, for every set of arithmetical sentences of finite logical complexity, L_{PA} can define a formula that represents truth locally. However, it cannot amalgamate these different truth predicates in a single formula holding universally for every L_{PA} sentence.

This result, which seems to testify the weakness of L_{PA}, apparently contrasts with the following positive result: although the set of true arithmetical sentences cannot be defined in L_{PA}, we can define the set that has such a set as unique member.

2.3 Theorem:

The class {T} the only member of which is the set T, which is the set (of codes) of true sentences in the standard model, is arithmetically definable.

[55] Boolos, Burgess and Jeffrey 2007.

INTRODUCING THE TRUTH PREDICATE[56]

Since truth is not representable using just PA, we have no choice: we have to add a new predicate and enrich PA. We do that in two steps: first, we introduce a new predicate constant into the language L_{PA}, with new rules of well-formation; second, we enrich PA with axioms and/or rules governing such a predicate constant, so that we can take it to be a truth predicate. The hard problem is clearly the second. Since the choice of what axioms and rules should be added requires a set of principles doing justice to truth. The simplest idea we can start from is that of Tarskian sentences. In the light of Tarski's adequacy condition Tarskian biconditionals must be consequences of every good definition of truth. Thus, we can simply try to build a theory in which the axioms are directly modelled on Tarskian biconditionals.

Let L_T be the language $L \cup \{T\} \cup \{c_i \mid i \in \omega\}$, where L is a first order language defined in the usual way, the symbol "T" is a new one-place predicate constant that represents our truth predicate and, possibly, an infinite numerable set of new individual constants. The intention is that the new individual constants give a name for every sentence of L_T. We adopt the convention according to which we construct such names putting sentences between quotation marks. The well-formation rules for those symbols are the usual ones. In the particular case we are considering, the language L is the language of PA, $L_{PA.}$ This fact allows us, thanks to the arithmetization of syntax, to avoid the

[56] The axiomatic theories of truth I discuss in this book are often referred to under other names. For example DT and its variants are often also named after the label TB (for Tarskian Biconditionals), and T(PA) and its variants are often also named after the label CT (for compositional truth). I hope that this variety of terminology does not cause confusion in the reader.

addition of new special individual constants. Accordingly, I write a sentence between cornered brackets $\ulcorner\varphi\urcorner$ to speak of the Gödel numbers of the corresponding sentence φ. This is an example of the utility of having a rich base theory like PA. Note, however, that if our base theory is not so strong the addition is mandatory. I use normal quotation marks to obtain a name of a sentence not in L_T, or when PA is not available, or outside a purely arithmetical context.

Once the language is enriched, we can pass to consider the specific axioms governing the behaviour of the new predicate "T". The theory[57] DT| (from Disquotational Truth)[58] consists in the axioms of PA in L_{PA}, with the addition of every sentence in L_T with the form:

TS: $\quad T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$

where φ is a sentence in L_T such that "T" does not occur in it . The restrictive clause on T-sentences is clearly a drastic measure needed to avoid that the theory contains paradoxical sentences like the liar. If we let φ be any sentence in L_T, then DT| would be inconsistent. Avoiding any occurrence of the truth predicate might seem to be a too severe restriction, but its rough simplicity allows it to avoid difficult and distracting topics at this point.

In formulating DT| we require the original axioms of PA to be in L_{PA}. This could seem obvious, but it is not. Instead, there are reasons to think that this is not the best way to add

[57] We take DT| (and the other axiomatic theories of truth) to include the base theory PA, because, generally, this is quite useful and unproblematic. Sometimes however (see infra Chapter Five) we refer to the mere truth theoretic part of the theory. In that case, for instance, DT| is the theory in L_T (as defined above with the addition of an infinite number individual constants) characterized only by T-sentences without the base theory PA. Clearly, in that case it is not necessary to distinguish between theories where full induction is or is not permitted: DT| and DT coincide.

[58] DT and its variants are often also named after the label TB, from Tarskian Biconditionals.

a new theory to a base theory like PA. The reason is that PA has a schematic axiom, the induction schema:

$$F(0) \wedge \forall x\,(F(x) \rightarrow F(Sx)) \rightarrow \forall x\,F(x)$$

where F(x), with x free, is any formula of the language in which the theory is formulated. It is worth reminding that the schema is just an expedient to mimick, in a first order way, what could be adequately formulated with the resources of a second order language. The idea of the induction principle is that it holds for every subset of natural numbers. However, what subsets are first order definable depends on the resources of the language we use. In fact, every formula $\alpha(x)$ in L_{PA} with x as a unique free variable defines a subset of natural numbers. If we enrich our language with new symbols, new formulas that define new subsets become available. If these new formulas can enter the induction schema, new instances of such a schema are obtained. So, when we add the predicate "T" we have to take a stand: either we allow this new symbol (via the new formulas formed with it) to enter the schema, so that new instances are obtained, or we stick with the old instances only (in this case the induction axiom is treated as a mere list).

Although a fully extended induction is not as innocent as it might seem and it often makes a big difference[59], it is important to notice that, reflecting on the nature of the schema, the enrichment seems to be not only natural but also an improvement of the schema. It enables us to define a bigger number of subsets of natural numbers. Allowing full induction in L_T, we obtain a new theory, DT, whose axioms are the axioms of PA in L_T with the addition of the sentences in L_T with the form:

$$\text{TS: } T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi$$

[59] For instances, see below the differences between T(PA)| and T(PA).

where φ is a sentence in L_T such that "T" does not occur in it.

Adopting T-sentences we have obtained two truth theories: DT|, with induction restricted to the language L_{PA}, and DT, with full induction in the language L_T. Although the distinction between DT| and DT is generally important, for the sake of simplicity, we mostly focus on DT now.

THE WEAKNESS OF DT

2.3 Proposition:

for any φ in L_{PA},

$$DT \vdash T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \neg \varphi \urcorner)$$

The proof is simple: since for any sentence φ in L_{PA} we can prove by simple logic that for any φ, $\varphi \vee \neg\varphi$ using two axioms of DT, $(T(\ulcorner \varphi \urcorner) \leftrightarrow \varphi)$ and $(T(\ulcorner \neg \varphi \urcorner) \leftrightarrow \neg\varphi)$, we get $T(\ulcorner \varphi \urcorner) \vee T(\ulcorner \neg \varphi \urcorner)$.

This result is especially interesting if compared to the following. Let Neg(x,y) be the L_{PA}-formula that says that x is the Gödelian of the negation of the formula with Gödelian y. Let, moreover, $Sent_{PA}(x)$ be the formula that says that the formula with Gödelian x is a sentence of L_{PA}. These two simple properties are recursive and therefore they are representable in PA.

2.4 Proposition:

$$DT \nvdash \forall x[SentPA(x) \rightarrow (T(x) \vee \exists y\ (Neg(y,x) \wedge T(y)))]$$

2.3 and 2.4 states that DT can prove the truth of every instance of a logic principle (in our case the law of excluded

middle, but it is easy to get a similar result for any other logical principle), but it cannot amalgamate these instances in a single generalization. This fact can be generalized stating that DT can prove only finite generalizations.

2.5 Proposition:

If DT proves a generalization of the form $\forall x(\alpha(x) \rightarrow T(x))$, where $\alpha(x)$ is an L_{PA}-formula with x free, then PA can prove that there are at most n objects satisfying $\alpha(x)$. PA proves $\exists n x \alpha(x)$ for some particular n.

(where "$\exists n x$" is an abbreviation of "$\exists_1 \ldots \exists_n (x_1, \ldots, x_n)$").

Similar results lead quickly to another result: DT cannot prove L_{PA}-sentences not already provable in PA alone.

2.6 Theorem:

For any sentence φ in L_{PA},
if $DT \vdash \varphi$, then $PA \vdash \varphi$

Proof:

We show how every proof in DT of a sentence φ in L_{PA} can be transformed into a proof of φ in PA. Suppose we have a proof Δ of a sentence φ in L_{PA}. Since the number of sentences occurring in this proof is finite, there must be an upper bound to the complexity of sentences occurring in Δ. Suppose the most complex sentence in Δ is Σ_n. We know that there exists an arithmetical partial truth predicate for Σ_n-sentences, and that PA can prove the Tarskian biconditionals for this partial truth predicate for all Σ_n-sentences. So we replace all occurrences of the primitive truth predicate "T" in Δ by occurrences of the partial truth predicate for Σ_n-sentences. This gives us a modified proof Δ', which is a

proof in PA. Moreover, the conclusion of Δ' remains φ, since φ does not contain any occurrences of "T".

THE TARSKIAN THEORY

The results above show that DT, the simple theory built on T-sentences, is a very weak theory. Tarski[60]already recognised and underlined this weakness, and for such a reason he searched for a suitable stronger theory. The result in 2.4, for example, is a clear case of what we can consider an inadequacy of a truth theory. Intuitively, we would like our theory to be able to prove generalizations about truth and involving only laws of logic. If we accept classical logic, we know, for example, that

$$\text{ExMiddle: } \forall x[\text{Sent}_{PA}(x) \rightarrow (T(x) \vee \exists y(\text{Neg}(y,x) \wedge T(y)))]$$

is true. Therefore, we might expect such a principle to follow from logic and a theory of truth. Otherwise, there would be something we know about truth, but such that our theory cannot prove. To get a proof of a generalization like ExMiddle we need to intervene on proposition 2.5: we must enable the theory to prove infinite generalizations. A possible way to do that is to keep following Tarski and get axioms for the truth predicate not from Tarskian biconditionals -namely from the mere criterion of material adequacy-, but from the clauses of the Tarskian definition of truth. In this way, what is yielded is a theory where the language is the same of DT, but the specific axioms governing the truth predicate, and added to those of PA, are now the following:

$$1.\ \forall x[\text{AtomSent}_{PA}(x) \rightarrow (T(x) \leftrightarrow T_0(x))]$$

[60] Tarski does not propose an axiomatic theory of truth, he proposes a definition. The relevant kind of considerations, though, is the same. See Tarski 1956.

2. $\forall x\ [Sent_{PA}(x) \rightarrow (\neg T(x) \leftrightarrow \exists y\ (Neg(y,x) \wedge Ty))]$

3. $\forall x \forall y\ [Sent_{PA}(x) \wedge SentPA(y) \rightarrow (T(x) \wedge T(y) \leftrightarrow \exists z(Conj(x,y,z) \wedge (T(z))]$

4. $\forall x \forall z \forall y\ [Form_{PA}(x) \wedge Free(z,x) \rightarrow (\exists w(Sub(x, z, y, w) \wedge T(w) \leftrightarrow \exists u Gen(u,z,x) \wedge (T(u))]$

Where "$AtomSent_{PA}(x)$" is the formula that says that x is (a code of) an atomic sentence in L_{PA}; Sub(x, z, y, w) is the formula that says that the formula (coded by) w is the formula we get by substituting the occurrences of the free variable z in the formula x with the variable y; Gen(u,z,x) says that the formula u is the formula obtained by binding with a universal quantifier the free occurrences of the variable z in the formula x; "T_0" in 1. is the local truth predicate for atomic sentences that is definable in PA. These axioms can be made more easily readable in the more intuitive following form:

1b. for any $AtomSent_{PA}$ $\varphi \in L_{PA}$: $(T(\ulcorner\varphi\urcorner)) \leftrightarrow (T_0(\ulcorner\varphi\urcorner))$;

2b. for any $Sent_{PA}$ $\varphi \in L_{PA}$: $(T(\ulcorner\neg\ \varphi\urcorner)) \leftrightarrow (\neg T(\ulcorner\varphi\urcorner))$;

3b. for any $Sent_{PA}$ $\varphi, \psi \in L_{PA}$: $(T(\ulcorner\varphi \wedge \urcorner)) \leftrightarrow (T(\ulcorner\varphi\urcorner) \wedge T(\ulcorner\psi\urcorner))$;

4b[61]. for any $Form_{PA}$ $\varphi(x) \in L_{PA}$: $(T\ulcorner\forall x\varphi(x)\urcorner) \leftrightarrow (\forall n T\ulcorner\varphi(\mathbf{n})\urcorner)$.

Also with such a Tarskian axiomatization, we face the same dilemma found met with DT: should we allow full induction in the new language or not? Depending on the choice, two different theories can be obtained[62]: T(PA)|, with

[61] Since the language contains a name for every natural number, truth does not require here a detour through satisfaction.

[62] T(PA) and its variants are often also named after the label CT, from

induction restricted to L_{PA} and T(PA), with full induction in L_T. The two theories have some notable differences. The first thing to notice is that both T(PA)| and T(PA) can overcome some weaknesses of DT: no analogous of the proposition 2.5 holds for them, since both theories can prove infinite generalizations.

2.6 Proposition:

Both T(PA)| and T(PA) prove
ExMiddle: $\forall x[SentPA(x) \rightarrow (T(x) \vee \exists y(Neg(y,x) \wedge T(y)))]$

A similar result holds analogously for the generalization of any logical law and not only for the excluded middle. Although both T(PA)| and T(PA) can prove infinite generalizations they do not prove the same generalizations: there are generalizations that T(PA) can prove but T(PA)| cannot. In fact, T(PA) does not prove only logical laws, like generalizations involving truth and logic, but it can also prove generalizations about the base theory PA. This seems to be a further evidence of adequacy with respect to truth theories based on simple T-sentences.

2.7 Proposition:

T(PA) proves:
T-Ax_{PA}: $\forall x[Ax_{PA}(x) \rightarrow T(x)]$
T-Inf_{PA}: $\forall x \forall y \forall z[(Inf_{PA}(x,y,z) \wedge T(x) \wedge T(y)) \rightarrow T(z)]$

Where $Ax_{PA}(x)$ is the L_{PA}-formula that represents in PA that the sentence with Gödelian x is an axiom of PA, and $Inf_{PA}(x,y,z)$ is the L_{PA}-formula that represents in PA that the sentence with Gödelian z is the result of an application of a

Compositional Truth.

rule of inference of PA to the formulas with Gödelian x and y. Obviously, such a result does not hold in the case of DT| or DT, since T-Ax_{PA} and T-Inf_{PA} are infinite generalizations, so they are beyond the strength of such theories.

A notable result is that T(PA) can prove that all theorems of PA are true.

2.8 Theorem:

T(PA) proves
T-$Teor_{PA}$: $\forall x[Prov_{PA}(x) \rightarrow T(x)]$

Proof: (sketched)

The theorem is proved by carrying out, inside T(PA), an induction on the length of proofs in PA. We do one basic case as an example. T(PA) proves the T-sentence $\forall x \neg(Sx=0) \leftrightarrow T(\ulcorner \forall x \neg(Sx=0) \urcorner)$ (see proposition 2.15). Combining this with the axiom of PA $\forall x \neg(Sx=0)$, T(PA) proves that $T(\ulcorner \forall x \neg(Sx=0) \urcorner)$ (similarly for the other axioms of PA). For the induction step, we need to show in T(PA) that

if $\exists x \exists z(Teor_{PA}(x) \wedge (z = \ulcorner \neg x \vee y \urcorner)) \wedge Teor_{PA}(z))$, then T(y).

By the inductive hypothesis, we have T(x) and T(z). By the truth axioms of T(PA) we infer from T(z) that $\neg T(x) \vee T(y)$. Combining this with T(x), we obtain T(y). So we can apply the principle of induction inside T(PA) to obtain $\forall x(Teor_{PA}(x) \rightarrow T(x))$.

Moving from the premises that all axioms of PA are true and that the rules of inference of PA preserve truth, T(PA) proves (with full induction) that everything PA proves is true. T(PA)|, instead, cannot prove T-$Teor_{PA}$ because the proof essentially needs formulas in L_T to appear into the induction schema: we need instances of the schema in L_T to get the conclusion. Since T(PA)| has restricted induction only, it

cannot prove the desired conclusion. Notice that both DT| and DT (even if DT has full induction in L_T!) cannot prove T-Teor$_{PA,}$ since they cannot prove infinite generalizations. DT| and DT are both unable to prove sentences in L_{PA}, unless those sentences are already provable in PA alone. Under this light, the difference between T(PA)| and T(PA) becomes very important. T(PA)| cannot prove new sentences in L_{PA}, whereas T(PA), thanks to the ability of proving T-Teor$_{PA}$, can.

2.9 Theorem[63]**:**

For any φ in L_{PA},
if T(PA)| ⊢ φ then PA ⊢ φ

2.11 Theorem:

T(PA) proves Con$_{PA}$, that is the sentence: ¬∃x(Prov$_{PA}$(x,⌈0=S0⌉)

Proof:

(Let us abbreviate ¬∃x(Prov$_{PA}$(x,y)) with ¬Prov$_{PA}$(y), and thus ¬∃x(Prov$_{PA}$(x,⌈0=S0⌉) with ¬(Prov$_{PA}$(⌈0=S0⌉)

1) T(PA) ⊢ ∀x(Prov$_{PA}$(x)⟶T(x)) (By Theorem 2.8)

2) T(PA) ⊢ Prov$_{PA}$(⌈0=S0⌉)⟶T(⌈0=S0⌉) (by instantiation of 1.)

3) T(PA) ⊢ T(⌈0=S0⌉)↔(0=S0) (by T-sentences)

4) T(PA) ⊢ ¬(0=S0) (because PA ⊢ ¬(0=S0) and PA is a subtheory of T(PA))

[63] The proof of this fact is not trivial at all. See Halbach 1999a.

5) $T(PA) \vdash \neg T(\ulcorner 0=S0 \urcorner)$ (from 4. and 3. by modus tollens)

6) $T(PA) \vdash \neg Prov_{PA}(\ulcorner 0=S0 \urcorner)$ (from 5. and 2.by modus tollens)

The sentence $\neg\exists x(Prov_{PA}(x,\ulcorner 0=S0 \urcorner)$ (namely, Con_{PA}) is in L_{PA} and, for the second Gödel's incompleteness theorem, it is not provable in PA alone. In fact, $\neg\exists x(Prov_{PA}(x,\ulcorner 0=S0 \urcorner)$ is the sentence that says that (a code of) a proof of the sentence 0=S0 does not exist in PA. In other words, Con_{PA} says that 0=S0 is not provable in PA. Since T(PA) proves ¬(0=S0), if it could also prove the negation of Con_{PA} it would be incoherent. Con_{PA}, then, is equivalent to the statement that PA is coherent but, for Gödel's theorem, PA cannot prove its coherence, then it cannot prove Con_{PA}. So there exists a sentence in L_{PA} that T(PA) proves but PA does not.

COMPARING THEORIES

The axiomatizations proposed so far have different characteristics, especially for the ability of proving infinite generalizations and new sentences in the language of the base theory. Such two aspects are actually strictly connected: the ability of T(PA) to prove a particular infinite generalization, $T\text{-}Teor_{PA}$, allows to prove a sentence that PA alone cannot prove. The theories that are not able to prove such a generalization (DT|, DT and T(PA)|) also fail to prove new sentences in L_{PA}.

Reflecting on the results above, another thing is worth noticing. These truth theories give give rise to a sort of hierarchy, from the weakest truth theory, DT|, to the strongest one, T(PA).

2.12 Proposition:

DT| is a proper subtheory of DT.

2.13 Proposition:

DT| is a proper subtheory of T(PA)|.

2.14 Proposition:

T(PA)| is a proper subtheory of T(PA).

2.15 Proposition:

DT is a proper subtheory T(PA).

It can be useful to sum up the main differences among such theories of truth on the base on their proof strength and on their impact on the base theory PA.

	(Some) infinite generalizations	**T-Teor$_{PA}$**	**New sentences in L_{PA}**
DT|	no	no	no
DT	no	no	no
T(PA)|	yes	no	no
T(PA)	yes	yes	yes

DEFLATIONISM AND THE AXIOMATIC THEORY DT

DT (and its restricted version DT|) is an axiomatic theory based only on T-sentences in the same spirit in which deflationism is a philosophical approach that maintains that T-sentences suffice to explain every fact involving truth. This shared idea allows to see DT as a formal counterpart of deflationism. Indeed, DT can be taken to be deflationism restricted into a formal axiomatic framework. Such identification is not immune from worries and some clarifications are in order. First of all, as we know, "deflationism" is more a title for a set of similar but different conceptions. It is worth clarifying which among these theories, and to what extent, can be represented by DT. Second, since DT treats the T-sentences as axioms for a new predicate, we have to verify that such an axiomatic approach is compatible with deflationism. Finally, we must consider paradoxes and the legitimacy of working with a restricted set of T-sentences in which the truth predicate is only applied to sentences in which the truth predicate does not occur.

The first point is strictly connected with the problem of truth bearers. Modern deflationism considers a range of possible truth bearers: sentence types, propositions and interpreted sentences are main options. Introducing the notion of base theory we have stressed that an axiomatic theory can fit a number of options concerning the nature of truth bearers. Consider, for instance, PA, which is the base theory we are working with and the most typical base theory. Since in PA the context is formal, we have not to handle demonstratives or context-dependent expressions. The lack of context-dependent expressions makes the distinction between types and tokens quite superfluous, at least for many practical purposes. Arithmetization of syntax

is then available. Thanks to the arithmetization of syntax we are then able to get a code from the symbols occurring in a sentence in such a way that a single code corresponds to every sentence, and, vice versa. Should a code $\ulcorner\varphi\urcorner$ be regarded as the name of the *sentence* it codes? The natural and immediate answer seems positive but this is not the only option. It is also possible to consider the coding process as outputting names for the *proposition* a sentence expresses. This tells us nothing about the nature of the proposition in question. The fact that it is possible to get a code does not mean anything about a supposed internal structure of that proposition. All we have to admit is that every sentence expresses a single proposition[64] and if two propositions are different, then the sentences expressing them are different too. Finally, if every sentence expresses a proposition, the contrary does not need to be true: not every proposition must be expressed by some sentence. A possible complication here is that we are dealing with a language as a pure syntactical entity, since it is not equipped with an interpretation yet. If a sentence is just a sequence of symbols, what do we mean by saying that such an uninterpreted sentence expresses a proposition? We can say that a sentence expresses a proposition at most if a model is also given. However, here we do not need to identify such propositions. We just need to know that for every interpretation of our language, suitable relations between sentences and propositions hold. Similar considerations can be put forward if truth bearers are identified with interpreted sentences. A code $\ulcorner\varphi\urcorner$ gives us a name of the interpreted sentence φ^{M} according to the model M. Hence, $\ulcorner\varphi\urcorner$ would be a name for a truth bearer only relative to a model. We hope that these rough observations are enough at least to show that also in an axiomatic approach we may

[64] Again, since we are in a formal framework this hypothesis is quite reasonable.

be able to make some sense of a philosophical approach based on propositions.

A second question concerns the relation between deflationism and an axiomatic treatment of the truth predicate. Under an axiomatic approach the notion of truth is treated as primitive, non definable and governed by axioms governing its behaviour. In the case of DT such axioms are the T-sentences. Considering T-sentences as axioms for truth is not, however, the only possible choice. According to Tarski the biconditionals are necessary constraints on a definition of truth formulated in a metalanguage, rather than axioms. Deflationists (in general) have not a distinction between object-language and metalanguage and they do not aim to give a definition of truth. Thus, T-sentences cannot be constraints on a definition. Rather, they are the very theory of truth. An axiomatic approach is thus natural to deflationism. Indeed, the goal of giving a definition of truth, which is the aim of a model theoretic approach like Kripke's, seems closer to the rivals of deflationism: substantialist theories of truth. A correspondence theory, for instance, searches for a definition of truth in terms of other notions such as correspondence to facts. The deflationist (and primitivist) idea that truth is undefinable naturally leads to an axiomatic approach as the most comfortable and consonant framework[65]. A final remark concerns the equivalence involved in the T-sentences. Different deflationary views can favour different readings of that equivalence: material, intensional, cognitive, analytic… In the present context we stick with material equivalence. Apart from simplicity, choosing a material biconditional has the advantage of being a weak option that is acceptable even to strongest views.

The last issue is the liar paradox. The problem of

[65] For more on this topic see Halbach 1999b.

paradoxes is an enormous problem for a theory of truth[66], and for a deflationist it seems even more serious. On the one hand the austerity of the theory makes the room for manoeuvre very narrow, on the other hand that an inconsistency looms over T-sentences seems to reveal that, after all, such principles are not completely innocent as the deflationists pretend. This is a strong hit against the legitimacy of a deflationary proposal. Moreover, the need for some measure to tame the paradoxes can easily force the conclusion that the T-sentences do not exhaust what must be said about truth. However, for the moment, we put this deep and general worries aside and limit ourselves to a safe set of T-sentences. In favour of this radical choice[67] we can point out that the set of T-sentences in DT (and in DT| as well) seems to be a subset of any candidate for set of T-sentences characterising a deflationary theory. In other words, although it is not clear what T-sentences should be contained in a deflationary theory, at least these ones[68] should be contained. DT (or DT| at least) seems the minimal formal counterpart of every conception of truth and, a-fortiori, of deflationism.

[66] See Simmons 1999 and Beall e Armour-Garb 2006.

[67] We only consider T-sentences where truth is ascribed to sentences in which the truth predicate does not occur.

[68] Even if grounded T-sentences (in a Kripkian sense) were preferred, our set would be a subset of such grounded T-sentences.

PART TWO

CHAPTER THREE
DEFLATIONISM AND CONSERVATIVENESS

FORMAL AND PHILOSOPHICAL THEORIES OF TRUTH

Both the approaches to truth met in the previous chapters - the philosophical approach and the formal - are subjects of big, if not huge literature. However, quite curiously, such literatures developed and kept developing rather independently from one another. Hardly the bibliography cited in an article in the philosophical field includes references to works in the logico-mathematical field and vice versa. Such a situation, which only in recent times has very slightly begun to change, is a serious reason for disappointment. Philosophical conceptions try to work out global visions locating truth in a general account of reality. However, at the same time, they often do not provide enough technical details to permit a sufficiently precise application and evaluation of the proposals. They remain vague or imprecise views in a number of cases. On the other hand, the theories developed in the logico-mathematical field offer a great deal of results, technically detailed and far from trivial, which make the theories liable to be precisely tested in different contexts. This approach, however, can hardly give answers to general questions about truth and its

place in inquiry. The general significance of such theories, if any, remains usually ignored. Such a situation, no matter how disappointing it is, does not seem to be a mere accident. Indeed, reading the two previous chapters, such a state of the art seems just the natural outcome. After all, the two approaches focus on deeply different topics, they ask different questions, and they employ different tools, with a few points in common if any.

Here is where deflationism can do an important job. Since deflationism holds that T-sentences, taken as principles of truth, are able to explain every fact about truth, it is quite easy to turn such a philosophical proposal into a formal theory and to investigate it with formal techniques, as argued at the end of the previous section. The argument from conservativeness, on which we are going to turn next, is a shining example of how profitable this interaction can be. By using technical notions, made available by the formal treatment of deflationism, we show that philosophical progress can be made to solve subtle and confused metaphysical questions. At the same time, philosophical considerations can shed light and guide the elaboration and the formal investigation of axiomatic theories. Although extended use of logico-mathematical tools is not new in analytical philosophy, this can be seen as a further confirmation of the power and usefulness of such a method.

THE ORIGIN OF CONSERVATIVENESS: HILBERT METAMATHEMATICS

In the second chapter we have seen that some axiomatic theories of truth (DT|, DT, T(PA)|) cannot prove sentences in the language of the base theory PA that are not already provable in PA alone. They cannot prove new sentences in

L_{PA}. This property, that can be generalized over the simple case of PA, is called *conservativeness* and it can be given two different forms: a proof-theoretic and a model-theoretic one.

1. **Proof-theoretic conservativeness***:*

A theory T in a language L_T is proof-theoretically (or deductively) conservative over a base theory B in the language L_B, if, for every sentence φ in L_B,

if $T \cup B \vdash \varphi$

then $B \vdash \varphi$.

Analogously, T ∪ B is said to be a proof-theoretic conservative extension of the base theory B.

2. **Model-theoretic conservativeness***:*

A theory T in a language L_T is model-theoretically (or semantically) conservative over a base theory B in the language L_B, if for every sentence φ in L_B,

if $T \cup B \vDash \varphi$

then $B \vDash \varphi$.

Analogously, T ∪ B is said to be a model-theoretic conservative extension of the base theory B.

The completeness theorem[69] for first order logic ensures that, at first order, the two definitions above are extensionally equivalent. However, keeping them distinguished in mind is important, because in some contexts this makes and it will make a big difference. Now we stick with first order logic so that we can use the name "conservativeness" to refer to this property in general without having to specify one of the formulations.

[69] Taking correctness for granted.

Conservativeness is a technical notion that played a crucial role in the program of the foundation of mathematics of David Hilbert[70]. In 1900 Hilbert made a relation at the Second International Congress of Mathematicians in Paris, where he listed twenty-three open problems in mathematics, presenting what he considered the most important mathematical questions that should have been solved in the future. The second of these problems was a proof of the coherence of arithmetic. At that time a lot of paradoxes and contradictions were discovered in the most basic notions and inferences of mathematics. In particular, the set-theoretic notions introduced by Cantor seemingly led to several incoherences. Since by using the resources provided by set-theory it is possible to reconstruct the whole mathematics, mathematics itself appeared to be deeply infected by paradoxes. Such a situation was completely unacceptable: where, if even mathematics fails, could we find certainty and truth? Hilbert himself dealt with this problem, looking for a firm foundation of mathematics. His program, known under the label "formalism", comes in two steps. In the first step, arithmetic should be completely formalized, in order to have a formal system of arithmetic. Such a system should be free from any appeal to problematic notions like meaning or truth. In fact, the system should be treated as a mere syntactical entity. As a set of symbols manipulable simply for their symbolic form. The Hilbertian conception of a formal system as a set of strings of symbols (in which we have axioms, rules of inferences, etc…) is the base of what today is considered a formal axiomatic system. Even if we have not given a meaning to our formulas we can study syntactical properties and relations. We can investigate what strings can be derived in that system.

[70] About Hilbert's program see for instance Feferman 1998 and Raatikainen 2003.

Then there is the second step of Hilbert's program: once we have a formal reduction of arithmetic, a proof of coherence of arithmetic should be provided by proving the coherence of the pure formal system yielded in the first step. A translation in a formal system is necessary to get an effective control of the procedure by which we get mathematical proofs. Only if we have such effective control we can check that the axioms and rules do not lead to incoherencies. Here, however, comes a subtle point: even if we were able to prove that the formal system is coherent, who would ensure us that the very proof of coherence is reliable? What does ensure us that metamathematics is more reliable than simple mathematics? Hilbert's solution to this problem is finitism. Hilbert demands metamathematical proofs to use only strictly finitary resources. The idea is that as long as we restrict mathematical practice into the finite, we can be safe from paradoxes, since finitary mathematics is free from worries about coherence. Thus Hilbert aimed at proving the coherence of a formal system, formalizing the whole arithmetic, using only finitary and thus safe tools[71].

Later on, such a "coherence program" has been turned into a "conservativeness program". According to the conservativeness program, in the formal system of arithmetic the subsystem of finitary arithmetic (thought as having an independent and safe meaning) should be distinguished from the part of the system that uses infinitary and problematic concepts. The goal was that of showing that the addition of infinitary arithmetic to the finitary one yields a conservative extension of finitary arithmetic. According to conservativeness, this would have shown that every theorem of finitary arithmetic proved using infinitary tools could have been proved without any use of such infinitary

[71] It is often said that Gödel showed that Hilbert's program was impracticable. Actually things are more complicated, for example because Hilbert did not clarify what finitary mathematics precisely is.

tools as well. If so, the principles of infinitary mathematics would have revealed themselves to be a mere heuristics to simplify and to help mathematicians work. Any resort to problematic infinitary notions could have been avoided, perhaps at the price of making the proof very long or more complicated. Moreover, a conservativeness proof would have also given a coherence proof. In fact, if a theory T is conservative over a base theory B, which we know to be coherent, then also T must be coherent. The reason is that since an incoherent theory can prove any sentence, it would also prove new sentences in the base language. In Hilbert's program the base theory is finitary mathematics, the coherence of which is taken for granted. It is customary to claim that Hilbert's program collapsed when Gödel proved his famous incompleteness theorems. Be it as it may, beside the Hilbertian program, the notion of conservativeness has kept being used in proof theory, for instance as a general means to prove the coherence of theories.

PHILOSOPHICAL APPLICATIONS OF CONSERVATIVENESS

Born into meta-mathematics and formal studies, the notion of conservativeness has been applied also to philosophical issues. A first example of these applications, at the boundary of philosophy and logic, can be found in the work of Michael Dummett[72] on the problem of the definition and meaning of logical constants. One of the notions introduced by Dummett in order to give a criterion that can help us define and evaluate the rules governing the logical constants (solving for instance the problems of Prior's *Tonk* or giving a reply to conventionalism in logic) is the notion of *harmony*. The rules of introduction and

[72] Dummett 1978.

elimination of a logical constant (like those in the calculus of natural deduction) should be, according to Dummett, in harmony with each other. For Dummett, conservativeness can then be used to make the notion of harmony precise. In Dummett's view we can say that there is harmony between the two aspects of a use of a logical constant if the addition of such a constant with its rules yields a conservative extension. The addition of a new logical constant must not permit the derivation of new sentences in which the new constant does not occur.

A different and interesting application of the notion of conservativeness to metaphysical and ontological problems is proposed by Stephan Schiffer[73] . Schiffer aims at constructing a theory of *pleonastic entities* that could solve, or dissolve, a range of classical philosophical problems. In Schiffer's view, pleonastic entities are, for example, propositions, properties and events. These entities are characterized by the fact that they emerge from linguistic practices that reify the entities in question. An example of such practices is the inference from "the Pope is human" to "the Pope has the property of being human". These reifying practices are built on conceptually valid inferences (Schiffer calls them *something from nothing transformations*) where the conclusions refer to entities that are not referred to in the premises. Not every practice of this kind, however, yields an authentic pleonastic entity: this would lead to a proliferation without control of such entities. Pleonastic entities are characterized, instead, by the fact that the addition of the concepts that allows the relevant reifying practices yields conservative extensions of our previous theories of the world: they have no effect on the pre-existent causal order. Beside the specific aspects of Shiffer's proposal, what is worth noticing is how a technical notion like conservativeness is used to clarify ontological

[73] Schiffer 2003.

questions regarding the metaphysical nature of certain entities. Thanks to the notion of conservative extensions the alleged innocent nature of these entities is demonstrated.

Another philosophical application of conservativeness is the one proposed by Hartry Field in his classical *Science without Numbers*[74]. Field wants to provide a reply to the argument of Quine and Putnam[75] showing that modern science, physics in particular, commits to the existence of abstract entities like numbers. Very briefly, the argument, summed up by Quine, goes like that: "science would be hopelessly crippled without abstract objects. We quantify over them. In the harder sciences, numbers and other abstract objects bid fair to steal the show. Mathematics subsists on them and serious hard science without serious mathematics is hard to imagine[76]". The point is that since physical sciences quantify over mathematical entities, and, what exists is revealed by what we are willing to quantify over[77], accepting these sciences implies accepting the existence of those mathematical entities too. Field attacks this argument by arguing that quantification over mathematical entities is not really necessary to science: we could reconstruct the entire science without quantifying over abstract objects or numbers at all. His strategy makes an essential use of the notion of conservativeness in a Hilbertian spirit. While Hilbert meant to show that infinitary methods are superfluous to prove results of finitary arithmetic, Field wants to show that mathematics is superfluous to prove results in natural sciences, of which, as an example, Field takes a Newtonian gravitational theory. The project, again, comes in two steps: in the first step Field gives a translation

[74] Field 1980.
[75] Quine 1948, Putnam 1971.
[76] Quine 1995, p. 40.
[77] Or better, what exists is revealed by what our best theories of the world quantify over.

of the Newtonian theory in a nominalist language, and a nominalist formulation (call it N) of such a theory. In the second step, the addition of a mathematical theory M (that is taken to be a kind of set theory like ZFC) to N is shown to yield a conservative extension of N. In other words, let N be a mathematics-free theory of the natural world, M a mathematical theory and φ a sentence in the language of N (so it is mathematics-free too): if φ is provable by N ∪ M then φ is also provable by N. Intuitively, this means that the addition of mathematics is redundant with regard to what is described by the natural scientific theory N. Note that what seems to be relevant here is the chance of avoiding the use of mathematics to *prove* non mathematical sentences, so that what seems to matter here is proof-theoretic conservativeness.

There is an analogous argument aimed at showing that the success of sciences requires that truth has a substantial nature. This argument moves from the idea that it is the truth of our scientific theories that explains their success and reliability. But then, if truth has such a causal and explanatory role, it is a legitimate object of research and it must have a substantial nature, against what deflationists claim. Williams[78] has replied to this argument with a move that is very close to Field's, although in a less technical way. Williams argues that everything that is explained with truth could be explained without it. In such a form, Williams reveals a more redundantist inspiration than a modern deflationist one. Field, for instance, does not claim that everything that can be made with mathematics could be made without it: clearly without mathematics we could not do mathematics. In the same spirit modern deflationism does not demand that everything done with the truth predicate could be done without it. However, it seems natural to put

[78] Williams, M. 1986.

Williams' strategy in weaker and more acceptable terms by invoking conservativeness: the notion of deflationary truth is not substantial in natural science because it does not allow to prove any new scientific sentence that is not provable without it. This suggestion directly leads us to our main topic.

INSUBSTANTIALITY AND CONSERVATIVENESS

In the first chapter we saw that a crucial point for a deflationary proposal is represented by the thesis according to which truth is not a substantial property. This idea is the flag that deflationists wave to advocate the more attractive feature of deflationary views with respect to rival traditional substantialist conceptions. However, it is both curious and serious that deflationists have not been able to provide any good explanation of what such an insubstantiality is supposed to be. An unexpected help[79] has arrived from the critics: some authors - Stewart Shapiro, Jeffrey Ketland and Leon Horsten[80] - have argued that the insubstantiality of truth should be explained in terms of conservativeness. According to them, that deflationary truth is an insubstantial property means that the corresponding theory is conservative. Shapiro gives the following argument for the conclusion that conservativeness is essential to deflationism. Suppose that somebody, call him "Karl", correctly knows a theory B in a language that does not contain a truth predicate (we can just think this theory to be PA) and suppose he adds a truth predicate to his language and extends B to a theory B $\cup$ T adding axioms that govern the truth predicate. Assume that

[79] This is not very surprising, since it makes it possible to formulate a deep objection against deflationism.

[80] Shapiro 1998, Ketland 1999, Horsten 1995. Although Horsten's paper is often neglected, actually he was the first to explain deflationist insubstantiality thesis in terms of conservativeness.

T ⋃ B is not a conservative extension of B, then there is at least a sentence φ in the language of B, L_B, (which does not contain the truth predicate) that is a logical consequence of T ⋃ B but is not a logical consequence of B alone. So, it is logically possible that B is true but φ false, whereas it is not logically possible that B ⋃ T is true and φ false. Before Karl introduced axioms for truth he could have accepted both B and ¬φ but the addition of T adds enough content to rule out the falsity of φ. The principles of truth have not only innocent consequences, in the sense that they have not only "semantical" consequences involving the truth predicate. Instead, the addition of truth has substantial consequences and this reveals that truth itself is substantial. If we want to support the idea that truth lacks a substantial nature, on the contrary, the addition of the theory of truth T must yield a conservative extension of B. In model theoretic terms, if *M* is any model of B, then T must be added in such a way that *M* must be expandable[81] to a model *M'* of T ⋃ B. If truth is metaphysically thin every model of a theory without a truth predicate and axioms for it should be expandable to a model of a theory with this predicate and these axioms. This implies the conservativeness of the truth theory over the base theory. Truth should not make, for the world, any difference. We should have the same models for our base theories before and after the addition of a truth predicate. The addition could enrich our language and proof strength in general, but it must not make any difference with respect to the base models we want to speak about. If we came to know something new about these models, if there was a new sentence in the base language that is a logical consequence of T ⋃ B, then we ought to exclude certain models: truth would have an impact on reality. This impact is possible only admitting that truth has a substantial robust

[81] Shapiro 1998 (p. 497) speaks of "extension of the model", but he is using the term "extension" *loosely*. What he means is "expansion" (private communication).

nature. Shapiro[82] argues like that: if truth is not substantial, it should make no difference for extra-semantical facts, but if the extension was not conservative then it would make a difference, so if truth is not substantial it must be conservative. Shapiro's argument is proposed together with the encouraging fact that T-sentences (in some restricted form) do yield conservative extensions in some cases. For instance, DT is conservative over PA.

Ketland, differently from Shapiro, does not move from general considerations to draw the conclusion that deflationism is committed to conservativeness. Instead, he focuses on DT (and similar deflationary theories) proving its conservativeness over PA[83]. From the fact that DT is conservative Ketland deduces the following corollaries:

3.3 Corollary (the contentless principle):

No non semantical statements (in L_{PA}) follows (only[84]) from a deflationary theory of truth[85] unless it is a logical truth.

---

[82] "The result is general. Let Γ be any theory that can express its own syntax. Add a new predicate T to the language and to Γ one of the common theories whose consequences are the T-sentences. Call the new theory Γ'. Then any model of Γ can be extended to a model of Γ'. (…) It follows that Γ' is a conservative extension of Γ". Shapiro 1998, p. 509.

[83] Two points are worth noticing in the conservativeness proof of Ketland: the first is that in the model-theoretic proof, he uses expandability of models. This will be very important later (see infra Chapter Six). The second is that Ketland's proof does not hold for any theory, for example it does not hold for the empty base theory (see infra Chapter Five).

[84] Here Ketland identifies a theory of truth with the set of pure truth theoretic axioms (or rules) for truth. He does not take a base theory like PA to be a part of the theory of truth. In this way his point is completely general. For complications, however, see infra Chapter Five.

[85] By "deflationary theory of truth" we mean here an axiomatic theory based on T-sentences like DT.

3.4 Corollary (the irrefutability principle):

No non semantical contingent statements (in L_{PA}) could refute a deflationary theory of truth.

3.5 Corollary (the consistency principle):

A deflationary theory of truth is consistent with any consistent non semantic theory (in L_{PA}).

Ketland emphasises that such corollaries show a kind of analyticity and contentless that deflationary theories should arguably exhibit: adding truth we do not add anything substantial. According to Ketland, these metalogical properties can clarify and explain the alleged insubstantiality and redundancy[86] of deflationary truth, so that Ketland concludes: "if truth is not substantial - as deflationists claim - *then* the theory of truth *should* be conservative. Roughly: *non-substantiality ≡ conservativeness*"[87]. Ketland, hence, moves from considerations about facts regarding DT, or similar theories, to conclude that such facts are essentially bound to fundamental principles of deflationism.

Details aside, the main thesis of Shapiro and Ketland is the same: the insubstantiality of truth that deflationists have in mind has to be explained in terms of conservativeness. On the one hand substantiality is argued to be bound to conservativeness (Shapiro's argument) so that conservativeness is a necessary condition; on the other hand conservativeness exhibits features that fit with important aspects of insubstantiality (see the corollaries),

[86] In his presentation of deflationary theories Ketland probably assimilates too much modern deflationist positions to redundantism. If analogies are important, differences are important too. Modern deflationism cannot be reduced to redundantism.

[87] Ketland 1999, p. 79.

so that conservativeness seems a sufficient condition too. Moreover, the fact that a formal counterpart of deflationism like DT is really conservative over PA, confirms the viability of this route.

At this point it is undeniable that the idea of using the technical notion of conservativeness to clarify the concept of not substantiality seems a good idea. Such a combination is not only natural; it seems also well motivated. Moreover, the elegant and precise explanation we obtain can enlighten a lot of puzzling claims typical of deflationism. Such a positive evaluation, in the case of a very critical point for which other satisfactory explanations are hardly available, means that the resort to conservativeness is an opportunity we should not refuse easily. What a deflationist means by saying that truth has no substantial nature might be not completely clear or not fully exhausted by an explanation in terms of conservativeness, but conservativeness looks like a promising option worth exploring[88].

THE ARGUMENT FROM CONSERVATIVENESS

An adequate theory of truth, as every adequate theory, should be able to explain every fact involving the notion it theorizes about. In our case, it should be able to explain any

[88] An interesting point is worth noticing. If truth (like in traditional substantialist frameworks) was definable, it would be naturally conservative. In fact we could always eliminate it in favour of the definition. However, when a primitive notion is considered, it is not obvious that its addition is conservative, because we cannot explain it away with a definition. This makes the resort to the notion of conservativeness very interesting in this context. Truth is not analysable, but, at the same time, it does not add anything new to the base theory. An alternative way that probably deserves some consideration is the case of circular definitions in the spirit of revision theory. A notion circularly defined is both conservative and unavoidable.

fact involving truth. For instance, arguably, a truth theory should prove generalizations like:

Gen: for any sentence φ, ψ, if "φ" is true and "ψ" is true, the conjunction "φ and ψ" is true.

If our theory could not prove such a law, there would be something we know about truth, Gen, that our theory would be unable to explain. The theory would not be able to make sense of everything it should, so it would be an inadequate theory. Notice that the plausibility of Gen is evident in the moment we reflect on the fact that we accept every instance of it, so that we would expect to get the generalization too. Gen is true simply by logic and by what we know about the notion of truth.

Another claim our theory should be able to explain emerges when a theory of truth is added to a base theory. If we accept a base theory B (in the sense of being willing to assert its axioms, rules and, thus, theorems), we would expect our theory of truth to be able to prove the truth of B. In other words, our theory of truth should be able to show the equivalence between accepting B and accepting the truth of B. We can see this in the case of PA. We expect a theory of truth to prove generalizations like: "the axioms of PA are true", "the rules of inferences of PA preserve truth" and "everything PA proves is true". The first two (which are what we called "T-Ax_{PA}" and "T-Inf_{PA}") are cases of generalizations whose each single instance is accepted without hesitation[89]. Since we accept the truth of every axiom of PA and the correctness of every rule of inference, we expect also that everything we obtain by application of these rules to these axioms is still true. We thus want the conclusion that everything PA proves is true (namely what we called "T-$Teor_{PA}$"). Similar considerations are,

[89] For example, we get every instance of T-Ax_{PA} considering every axiom and the corresponding T-sentence.

prima facie, completely reasonable and lead us to state as adequacy clauses -so that we collectively indicate them as the *requirement of adequacy*- that a theory of truth should be able to prove generalizations like Gen, T-Ax_{PA}, T-Inf_{PA} and $\text{T-Teor}_{\text{PA}}$.

Now we can put together and apply to deflationism what we have just obtained. If a deflationary theory of truth is to be considered an adequate theory of truth, it must be able to meet the adequacy requirement on the one hand, and, given the explication of insubstantiality in terms of conservativeness, it must also be a conservative theory. Unfortunately here comes the problem: no theory of truth can satisfy both the adequacy and the conservativeness requirement. The reason is that if a theory of truth proves $\text{T-Teor}_{\text{PA}}$, then it is immediately able to prove Con_{PA}, the sentence in L_{PA} stating the coherence of PA, which cannot be proved in PA. Hence, there is a sentence φ in L_{PA} that can be proved by our theory of truth but such that cannot be proved using PA alone: conservativeness is lost. We have seen that this is what happens in the case of T(PA) (Theorem 2.11), but it is worth showing that what is needed is much less than the whole theory T(PA). Actually it is enough to add $\text{T-Teor}_{\text{PA}}$ to PA together with a weak formulation of the truth axioms[90] like DT| to obtain Con_{PA}. This fact shows that the assumption of the truth of the theorems of PA (namely $\text{T-Teor}_{\text{PA}}$) is the real responsible of the loss of conservativeness

3.4 Theorem:

$$\text{PA} \cup \text{DT|} \cup \{\text{T-Teor}_{\text{PA}}\} \vdash \text{Con}_{\text{PA}}$$

[90] Some axioms for truth are needed, otherwise the occurrence of "T" in $\text{T-Teor}_{\text{PA}}$ would be vacuous.

The proof is clear from the proof of the theorem 2.11. To be precise, only one T-sentence is needed: $T(\ulcorner 0=S0 \urcorner) \leftrightarrow (0=S0)$.

At this point deflationism is doomed to be an inadequate theory of truth. The argument can be summed up in a very straightforward way: a deflationary theory of truth must be a conservative theory, an adequate theory cannot be conservative, therefore a deflationary theory is inadequate.

If we compare the axiomatic theories met in the second chapter with the requirements of adequacy and conservativeness, we can see that only T(PA) satisfies the criterion of adequacy, and thus it fails to satisfy the conservativeness request. The other three theories, DT|, DT and T(PA)| are all conservative, so that they cannot prove all the expected generalizations. T(PA)| is the theory that comes closer to satisfy the adequacy requirement: it can prove infinite generalizations (on the contrary of DT| and DT) but being conservative on PA, T(PA)| is not able to prove all such generalizations. For example, it cannot prove $\text{T-Teor}_{\text{PA}}$.

THE GÖDELIAN SENTENCE

One of the recursive functions that we can represent in PA is the function $\text{Sub}_{\text{PA}}(m,n,p)$, which says that if m is the code of a formula φ in L_{PA} and n is the code of a variable x the function yields the code of the formula $\varphi(x/\mathbf{p})$ as value; which is the formula that we get substituting, in φ, the free occurrences of the variable x with **p**, which is the numeral of the number p. $\text{Sub}_{\text{PA}}(m,n,p)$ is the representation in PA of the operation of substitution.

Consider now the formula in L_{PA}:

$$G(y): \neg\exists x \text{Prov}_{\text{PA}}(x, (\text{Sub}_{\text{PA}}(y, \ulcorner y \urcorner, y))$$[91]

[91] In order to be, hopefully, more perspicuous, some details of the

such a formula can be read like "there is no x, such that x is the code of a proof in PA of the formula that we get when we substitute in the formula of code y, the free occurrences of the variable y with the numeral of the code of the same formula". Now let q be the code of the formula G(y), which is:

$$\ulcorner \neg \exists x Prov_{PA}(x,(Sub_{PA}(y, \ulcorner y \urcorner, y)) \urcorner = q.$$

let finally G be the sentence in L_{PA} that is obtained from G(y) by putting q in the place of the variable y:

$$G: \ \neg \exists x Prov_{PA}(x,(Sub_{PA}(q, \ulcorner y \urcorner, q))$$

G can be read as: there is no x such that x is the code of a proof in PA of the sentence whose code we get from $Sub_{PA}(q, \ulcorner y \urcorner, q)$. In other words: there is not a proof in PA of the sentence φ the code of which we get substituting, in G(y), the free occurrences of the variable whose code is $\ulcorner \varphi \urcorner$ with the numeral of the code of G(y). More shortly, we can say: the formula G, which we get by substitution of the free occurrences of the variable of code $\ulcorner \varphi \urcorner$ with the numeral of q in the formula q, is not provable in PA. Since the variable with code $\ulcorner \varphi \urcorner$ is y and the formula with code q is G(y), we have to substitute the variable y with the numeral of q in the formula $\neg \exists x Prov_{PA}(x,(Sub_{PA}(y, \ulcorner y \urcorner, y))$, so we get:

$$G': \ \neg \exists x Prov_{PA}(x,(Sub_{PA}(q, \ulcorner y \urcorner, q)).$$

G states that in PA a proof of the sentence G' does not exist. G', however, is just G, so we can say that G says[92] of itself that it is not provable in PA.

G is the famous Gödelian sentence that, according to

technical formulation are omitted, or changed. For example, the right distinction between numerals and numbers is neglected. Fully rigorous treatments can be found in textbooks such as Kaye 1991, Hájek and Pudlák 1993.

[92] Rigorously, what we have is a biconditional such that a formula G is true if and only if $\neg \exists x Prov_{PA}(x,(Sub_{PA}(q,\ulcorner y \urcorner,q))$.

the first incompleteness theorem, is neither provable nor refutable in PA. Moreover, G can be shown in PA to be equivalent to the sentence Con_{PA} which states the coherence of PA. This means that if we are able to prove Con_{PA}, we can prove G too. Hence, among our truth theories T(PA) is the only one that can prove G. This fact strengthens the previous non conservativeness result, since G is in L_{PA}.

This fact is, in Ketland view, another reason to consider T(PA) an adequate theory of truth, against conservative theories. We are actually able to "see" the truth of G (assumed the truth of PA), although G is not provable in PA. The (very informal) reasoning that is often sketched is the following:

G says of itself that it is not provable[93] *in PA; now, by the first Gödel incompleteness theorem we know that G is not really provable in PA*[94]*, therefore, after all, things are exactly as G says, so G is true.*

According to Ketland we can draw this conclusion thanks to our mastery of the concept of truth. Hence, though G is not provable in PA, it should be provable in an extension of PA, call it M(PA), which extends PA with truth axioms. Let us try to put the informal reasoning above in such a bigger (meta)theory. What we want is a deduction with the following form:

1. $PA \vdash G \leftrightarrow \neg Prov_{PA}(\ulcorner G \urcorner)$ - (where $\neg Prov_{PA}(\ulcorner G \urcorner)$ shortens $\neg \exists x Prov_{PA}(x,(Sub_{PA}(q, \ulcorner y \urcorner, q)))$).

2. $PA \nvdash G$

3. $M(PA) \vdash \neg Prov_{PA}(\ulcorner G \urcorner)$

4. $M(PA) \vdash G$

5. $MPA) \vdash T(\ulcorner G \urcorner)$

[93] Assumed that PA is ω-coherent.
[94] Nor is it refutable in PA.

The crucial step is represented here in the move from 2. to 3.; here is what we have informally read as: *G is not really provable in PA (*PA $\nvdash$ G *),* and from which we go on saying that, *after all, things are exactly as G says:* $\neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$*, then G is true.* Notice that such a step is not possible in PA. Step 2. says that PA does not prove G, while 3. says that M(PA) proves that G is not provable in PA. In one case we say that there is not a PA-proof of G, in the other case we say that there is a M(PA)-proof of $\neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$. If it was possible to pass from "PA $\nvdash$ G" to "PA $\vdash \neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$", then, since $\neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$ is just G, PA would prove G[95] [96]. What we want, thus, is exactly an extension of PA allowing this step. A metatheory that enables us to prove that if PA $\nvdash$ G, then we can conclude $\neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$. It is not possible to get such a deduction from a conservative theory (like DT|, DT or T(PA)|) since such a theory cannot prove G. The only theory of truth that can do that is T(PA). How we can pass from 2. to 3. in T(PA) can be seen below[97].

$$T(PA) \vdash G \leftrightarrow \neg \mathrm{Prov}_{PA}(\ulcorner G \urcorner)$$

$$T(PA) \vdash \forall x(\mathrm{Prov}_{PA}(x) \rightarrow T(x)) \qquad \text{(Theorem 2.11)}$$

$$T(PA) \vdash \mathrm{Prov}_{PA}(\ulcorner G \urcorner) \rightarrow T(\ulcorner G \urcorner) \qquad \text{(from 2. by instantiation.)}$$

[95] If PA proved G, then it would prove also $\mathrm{Prov}_{PA}(\ulcorner G \urcorner)$, and it would be incoherent. This is just the spirit of Gödel's proof.

[96] Here the fact that the function $\exists x \mathrm{Prov}_{PA}(x,y)$ is not recursive is crucial. If it was recursive, it would be representable in PA (and not only semi-representable), then if a sentence with code n is not provable in PA, this should be represented in PA by PA $\vdash \neg \exists x \mathrm{Prov}_{PA}(x,n)$. In such a way we could pass from 2. to 3. and PA would be incoherent. Since, however, $\exists x \mathrm{Prov}_{PA}(x,y)$ is not recursive this is not possible. Note that since it is semi-recursive, $\exists x \mathrm{Prov}_{PA}(x,y)$ is semi-representable, and it allows the "positive" step from PA $\vdash \varphi$, to PA $\vdash \exists x \mathrm{Prov}_{PA}(x,\ulcorner \varphi \urcorner)$, that is a property of the proof predicate.

[97] See Ketland 1999, p. 87.

$T(PA) \vdash \text{Prov}_{PA}(\ulcorner G \urcorner) \rightarrow G$ (from 3. by T-sentence)

$T(PA) \vdash \text{Prov}_{PA}(\ulcorner G \urcorner) \rightarrow \neg \text{Prov}_{PA}(\ulcorner G \urcorner)$ (from 4. and 1.)

$T(PA) \vdash \neg\neg \text{Prov}_{PA}(\ulcorner G \urcorner) \rightarrow \neg \text{Prov}_{PA}(\ulcorner G \urcorner)$ (from 5. by MTT)

$T(PA) \vdash \neg G \rightarrow G$ (from 6. and 1.)

$T(PA) \vdash G$ (from 7. by reductio)

A more straightforward and informal proof can also be given. We know that every axiom of PA is true, and that every rule of inference of PA preserves truth, thus every theorem of PA is true. This implies the sentence Con_{PA}, and since Con_{PA} is equivalent to G, we get G. This argument is usually known as the semantic argument, since the proof makes an essential use of the semantic predicate *par excellence*: the truth predicate. It is a similar argument that leads to the semantic interpretation of Gödel's theorem and that allows us to speak of the semantical incompleteness rather than the simple syntactic incompleteness, as in the original Gödelian proof. It is from similar considerations that it can be argued that truth transcends proof and that there are true sentences that are not provable. A deflationary theory, committed to conservativeness, cannot apparently make sense of such an ability to "see the truth" of G. The Gödelian sentence gives us another perspective under which a conservative theory does not seem adequate.

WHICH CONSERVATIVENESS?

Shapiro and Ketland move from slightly different considerations to conclude that a deflationary theory should be conservative. Their final formulation, however, is not completely satisfactory because it leaves some aspects to be explicitly addressed. Which conservativeness is at stake here? The semantical or the deductive one? Which kind of logical consequence is relevant? First order logic or a higher order logic? Finally, speaking of conservativeness simpliciter makes no sense: conservativeness always demands a base theory. On which base theory should our truth theory be conservative?

Ketland is not very explicit about what the relevant kind of conservativeness should be, but probably a first order proof theoretic conservativeness is what he has in mind. Since Shapiro is more explicit, we limit ourselves to consider Shapiro's arguments as a case study. First of all, an important specification is worth doing. Shapiro[98] does not use the argument from conservativeness to show the inadequacy of deflationism (as Ketland does). His aim, instead, is to show that deflationists have good reasons to embrace a notion[99] of logical consequence richer than first order. The reason is that a deflationist should be bound to conservativeness in some form, but given that if she sticks to first order logic she has no chance to get an adequate theory of truth, she should embrace a stronger logical consequence. "The only retort for the deflationist (short of surrender) is to insist on a separation between semantic/metaphysics matters and epistemic/proof matters. The deflationist must maintain that since truth is metaphysically thin we cannot prove anything adding a truth predicate *that was*

[98] As he argues in Shapiro 2002.

[99] Note that Shapiro advocates the opportunity for a stronger notion of logical consequence independently from deflationism.

not already implicit in the original subject".[100]A deflationist should insist that the truth predicate has an essential role in proving some sentences but that this does not imply anything about its substantial nature, if what it shows is something already implicit in the base theory. Truth, in this case, would not have any impact on reality and it would not have metaphysical substantiality. What matters for substantiality is not which sentences a theory can or cannot prove but which sentences were true (which sentences were logical consequence) before our addition of principles for truth. If semantical conservativeness holds, nothing that was not already there would have been obtained by truth. The idea is that if semantic conservativeness is respected, then deflationists have opened a possible way out.

Now let see how a higher logic makes such a way out viable. If first order logic is adopted, deductive and semantic conservativeness coincide and there is no way out. Such a way out is immediately available, however, if we abandon first order logic for a richer logic instead. Shapiro considers a number of cases: second order, analytic consequence and substitutional consequence. We focus just on the second order case to clarify the point. What we should do[101] is turning the induction schema of PA into a single second order axiom:

$$\forall X[(X0 \wedge \forall x\, (Xx \rightarrow XSx)) \rightarrow \forall x Xx]$$

where the deductive system for the second order contains also a comprehension schema:

$$\exists X \forall x\, (Xx \leftrightarrow \alpha(x))$$

in which instances for each formula α not containing X free are obtained[102]. In that system (the second order

[100] Shapiro 1998, p. 504, (italics added).
[101] I follow Shapiro 1998, p. 508.
[102] The other necessary modifications for passing from a first order to

version of PA, called "PA_2") when new symbols are added to the language new formulas automatically go in the comprehension schema and the extension obtained by these formulas gives new instances of the induction axiom. In such a system Con_{PA2} and G_2 become automatically provable the moment truth axioms are added[103]. The crucial point, however, is that those sentences are already true in the unique model of PA_2, so that they are semantically implied by PA_2. In the second order case provability has not the same extension of logical consequence; the former is a proper subset of the latter instead. This means that Con_{PA2} and G_2 are not theorems of PA_2 but they are (second order) logical consequences of PA_2. Thus, a truth theory can prove Con_{PA2} and G_2, without implying that the truth involved is substantial. If semantic conservativeness is the kind of conservativeness that matters for insubstantiality, and a higher order logical consequence is adopted, then Con_{PA2} and G_2 are already consequences of PA_2[104].

Apart from being a first possible deflationist reply to the argument from conservativeness, and being a specification of what is involved in the choice of a particular logic, these considerations clarify what kind of conservativeness seems important for the argument: semantical conservativeness. According to Shapiro's argument this does not mean that deductive conservativeness could never be a problem for a deflationist. For example, the role that truth might have in the explanation of the success of scientific theories[105], mentioned above, seems to bring the argument back in the direction of epistemology and proof-theoretic conservativeness. Similarly to Field's strategy to avoid

---

a second order version of PA are taken for granted.

[103] Shapiro 1998, p. 508. See also Shapiro 1991, Chapter Five.

[104] In fact PA_2 has only isomorphic models, so that we are allowed, in this sense, to talk about the unique model of PA_2.

[105] Putnam 1971.

ontological commitments to mathematical abstract entities, the relevant sense of conservativeness seems the deductive one. It is the deductive role of mathematics and truth in scientific inquiry what forces a commitment to the existence of mathematical entities. The point can also be put in terms of PA and the Gödelian sentence G. Suppose that a teacher of logic states that G is true and that a student asks for an explanation. The student, in this scenario, believes in the truth of G (perhaps he just trusts him) but he wants an explanation of why G is true. The teacher then uses the semantic argument: the axioms of PA are true, the rules preserve truth, therefore the theorems of PA are true. Hence 0=S0 cannot be a theorem of PA. Since G and Con_{PA} are equivalent, finally, also G must be true. This informal explanation is an explanation of why G and Con_{PA} are true and it is a case of deductive non conservativeness. If an explicative role is also a mark of robustness, deductive conservativeness cannot be put aside. Shapiro notices that in this case a deflationist cannot avoid the problem resorting to semantic conservativeness by claiming that G and Con_{PA} are logical consequences of the base theory. Indeed, it is this fact that must be explained and that is explained in terms of truth[106]. Shapiro, however, concludes that since there is no general consensus on the notion of explanation, it is quite hard to give a final evaluation. Therefore, although Shapiro argues in favour of the relevance of semantic conservativeness, he does not exclude in principle some possible roles for the proof-theoretic one.

While I agree with the dismissal of proof-theoretic conservativeness, Shapiro's considerations do not

[106] Shapiro suggests that a possible way out for a deflationist would be to argue that the only problematic explicative uses of the notion of truth are those rising in empirical and causal contexts. Such a solution, however, leaves unanswered the question of why certain explicatory uses commit to metaphysical robustness whereas others do not.

seem fully on point. First of all, Field's argument for the redundancy of mathematics is rather different from the one for the conservativeness of the notion of truth. In the former case what is advocated is the possibility of accepting a certain base theory without accepting the existence of certain entities. In this sense if the resort to mathematics is necessary for the deduction of new knowledge, we must use mathematical sentences quantifying over abstract entities committing ourselves to their existence. The case of truth is different. Deflationism in its modern formulation does not deny that truth exists but only that it is a substantial property. An argument showing that if truth is used in explanation then it has a robust nature is needed. Certainly a problem is still here for those deflationists who think that truth is redundant or it is not a property at all. If truth is essential then the idea that truth can be eliminated without expressive loss is untenable. For such a reason Williams[107] replied that truth is not really essential in explanations and that everything explained with truth could be explained without. In modern deflationism, however, crude redundantism does not seem very widespread, if held at all. We know that the truth predicate cannot be eliminated without expressive loss: the truth predicate allows the expression such as generalizations, blind ascriptions and so forth, that we could not express without. A truth predicate, then, enriches our language and theories. If we suppose that some of those expressions occurred essentially in a derivation of some new sentence, the resort to truth would be indispensable. The point we should focus on is whether such an indispensable resort commits to the substantiality of truth. Beware: in the case of semantic non conservativeness we do have such an argument, but here we are considering deductive non conservativeness. Perhaps, we could say

[107] Williams, M. 1986.

that truth when proof-theoretic conservativeness is lost, substantiality is revealed at an epistemic level: truth exhibits some epistemic/deductive robustness. But here "robustness" means just that truth has a crucial utility for us, and indeed the epistemic/deductive usefulness of truth is one of the main claims of deflationism. To insist that this epistemic usefulness is what proves the substantiality of truth does not help, because it is this very point that must be explained: why should epistemic usefulness imply a metaphysical substantiality of the property of truth if we have semantic conservativeness? In such a case it seems just as truth helps us speak about reality. Truth does not inflate the world. It is just an improvement of our linguistic/epistemic resources. Truth does not add anything to the world, it just makes us able to have a better grasp on it. In other words, mere epistemic robustness can always be explained in terms of epistemic usefulness, and a robust epistemic usefulness is completely acceptable to a deflationist. Thus, only if what we can prove is something not already implicit in the base theory we seem to have a problem, because it is in this case that truth imports substantial content. But this is exactly what is at stake in semantic conservativeness. In other words a possible role of truth in explanation/proof is not a mark of metaphysical robustness as long as semantical conservativeness holds.

Such reflections can be illustrated again with the case of PA and G. The student accepts the truth of G but he wants an explanation of why G is true. Shapiro notices that "it does not help for the deflationist to point out that Con and G are (after all) semantic o logical consequence of the original theory A because this was the fact that needed explanation[108]" This point is not completely clear. Shapiro

[108] Shapiro 1998, p. 505.

takes A to be a first order arithmetic theory[109], like PA, but neither Con_{PA} nor G are (first order) logical consequences of PA. After all, there are models of PA in which G is false. Thus, what Shapiro probably means is that Con_{PA} and G, given PA, are true in some intuitive sense. In other words, what is at stake here is the truth of G in the intended standard model $\mathbb{N}$: the student accepts that G is true in $\mathbb{N}$ but he asks for an explanation. If we assume that the student has in mind the standard model, however, we can give a way shorter explanation of Shapiro's and such an explanation does not involve truth at all. It suffices to point out the final part of the previous explanation: in the standard model it is not the case that 0=S0, so if $\mathbb{N}$ is the intended model of PA, PA cannot prove that 0=S0, thus it does not exist in PA a proof of 0=S0. This is exactly what Con_{PA} states. Since Con_{PA} can be shown to be equivalent to G, then also G is justified.

It might be objected that speaking of sentences holding in a model without the notion of truth is not possible, but this objection takes us back to the model-theoretic point of view and its relevance is not denied. Again, we can suspect that the notion of truth is hidden in the hypothesis of the coherence of PA. If the student was to ask for another explanation of this, we should remind him that the axioms of PA are true, that the rules of inferences preserve truth, so that every theorem of PA is true and PA is coherent. However, this explanation is not mandatory. If the student has already in mind the standard model $\mathbb{N}$ and he believes $\mathbb{N}$ to be a model of PA, this is enough to the conclusion that PA is coherent[110] [111]. We can still doubt whether the relation between the model $\mathbb{N}$ and PA could be clarified without using the notion of truth but, as above, this takes us back

[109] Shapiro 1998, p. 500.

[110] At least if the student knows the theorem of the existence of a model.

[111] Or we could even use Gentzen's proof.

to the model-theoretic perspective. Similar considerations hold also if Shapiro meant to refer with "A" to a second order arithmetic theory like PA_2.

However, it can be conceded, as Shapiro himself does, that the lack of deductive conservativeness can be a problem for some forms of deflationism, like redundantism, whereas modern deflationism, on the contrary, is hardly bothered. Since one of the flags of modern deflationism is the epistemic usefulness of the truth predicate, it is always possible, for such a modern deflationist, to find refuge in semantic conservativeness. In the words of Field (a deflationist that accepts the conservativeness requirement): "I shall call attention to one point to which every theorist of truth should agree: that by having a notion of truth we increase our expressive power in an important way[112]" and "Shapiro says that deflationists hold that truth is "metaphysically thin" (…) I am not sure what this mean, but one thing that is better not mean is that we cannot use it to express important things inexpressible or not easy expressible without it, or that we cannot use it to make commitments about matters not involving truth beyond this commitments which we could make without it, for it is a clear part of deflationist doctrine that truth is not metaphysically thin in that sense (We might put this by saying that everyone, deflationists or not, must agree that truth is not expressively thin)[113]".

The second point we have to clarify is what kind of logical consequence matters for the conservativeness requirement. Shapiro suggests that a deflationist should adopt a quite rich notion of logical consequence, because this move would allow the deflationist to escape the objection. In that way the argument itself does not seem to command any particular kind of logical consequence, since a deflationist can shift

[112] Field 1999, p. 533.
[113] Field 1999, p. 534.

to the logical consequence she prefers. In response, we just limit ourselves to some basic considerations. If a deflationist wants to argue in favour of a stronger logical consequence she should propose motivations that are independent from conservativeness. As Shapiro has widely showed, many of the proposed options[114] formalize a logical consequence that is not effective and not tractable under many respects. If deflationists were right to adopt one of them, they had better argue from strong independent reasons[115], otherwise their choice would seem to be not only *ad hoc* but even improper. Moreover, as Shapiro himself notes, the resort to a different logical consequence seems only an apparent way out: we would have hidden the robustness of truth into the robustness of logical consequence. Thus the problem would have been moved without being solved. The situation briefly is: if logic is thick, then truth is thin; if truth is thick, then logical consequence is thin. We should then probably agree with Halbach when he says: "This loophole, I suspect, is more a trap than a way out"[116]. For these reasons we shall concentrate on first order logic putting higher order logics aside, although it is not denied that a similar strategy could be a possible deflationist option, and proposal in that direction will be discussed again in due course.

The last issue we have to face in order to state the conservativeness requirement precisely concerns the identification of the base theory over which we must require conservativeness. If we focus on the reasons that led deflationism to be committed to conservativeness, only one option seems likely: our requirement must have a universal range. A deflationary theory should be conservative over any base theory. The argument, in fact, could be reconstructed

[114] But see Hyttinen and Sandu 2004.
[115] Clearly Shapiro, even if not a deflationist, thinks that he has such reasons.
[116] Halbach 2001a, p. 170.

for any arbitrary base theory. In other words, it suffices to find a single base theory on which the deflationary theory is not conservative to make the requirement not satisfied.

At this point we can put the previous points together - semantical conservativeness, first order logic, relevance of any base theory - and formulate the conservativeness requirement more precisely.

Conservativeness requirement:

if T is a deflationary theory of truth, for every base theory B, B $\cup$ T ought to be a conservative extension of the base theory B, where the relevant logical consequence is first order. More formally:

If T is a deflationary theory of truth in a language L_T, for every base theory B in a language L_B, and for every sentence φ in L_B,

if T $\cup$ B $\models \varphi$ then B $\models \varphi$.

Where " $\models$ " stands for the standard first order logical consequence relation.

CHAPTER FOUR
DEFLATIONIST REPLIES TO THE ARGUMENT FROM CONSERVATIVENESS

The argument from conservativeness aims at showing the inadequacy of deflationary theories of truth along the following lines:

> 1. a deflationary theory must respect the requirement of conservativeness;
>
> 2. an adequate theory of truth cannot respect such a requirement;
>
> 3. therefore, a deflationary theory cannot be an adequate theory of truth.

We have seen how and why a deflationist is committed to conservativeness by the thesis that truth lacks a substantial nature. We also know what is needed to consider a theory of truth adequate, and we know that this makes it impossible to respect a request of conservativeness. Some deflationists have tried to reply to this argument. The available options are:

> 1. to attack the first premise by denying that a deflationary theory of truth must be conservative;
>
> 2. to attack the requirement of adequacy. This can be made in two ways: 1a. By denying the validity of the requirement, or 2b. By showing that it is possible for a (deflationary) theory of truth to respect the

requirement by using principles external to a theory of truth.

3. to attack the second premise interpreting the relevant logical facts in a way showing that a deflationary theory is indeed able to satisfy both the conservativeness and the adequacy requirement.

All the above strategies have been proposed and all meet serious obstacles. The first option has been put forward by Volker Halbach[117] and it has the problem that, on the one hand, it charges the deflationist with the burden of explaining how it is possible for truth to be insubstantial if it is not conservative[118]; on the other hand, if we give up conservativeness we need an alternative explanation for the enigmatic insubstantiality of deflationary truth. It is not surprising, then, that deflationists have mostly omitted this option and the association between deflationism and conservativeness has been widely accepted. The second strategy is proposed in the strong form (2a) by Jody Azzouni[119] and, in some measure, by Dan Waxman[120], as we will see, although it is a straightforward reply, it seems to neglect a seemingly basic feature of truth, namely its reflective power. In the weak form (2b) the move has been adopted by Neil Tennant[121]. He tries to show that the resort to reflection principles allows the deflationist to obtain what is requested by the adequacy requirement. The attempt, however, is arguably not fully convincing. The last strategy, which tries to keep conservativeness and adequacy together, has been proposed by Hartry Field[122]. His proposal would

[117] Halbach 2001a.
[118] Cieslinski 2015 also supports this.
[119] Azzouni 1999.
[120] Waxman 2017. Waxman actually combines option 2a and option 3 in a disjunctive strategy.
[121] Tennant 2002.
[122] Field 1999.

make everyone happy. We would get a satisfactory theory of truth both for deflationists and their critics. Deflationism would be strengthened by this operation. Not only would have it replied to a deep objection, it would have also clarified a point that has been obscure and confused until now. Also Field's move, however, is problematic. Another response that follows the same line is that of Shapiro, who, as we saw in the previous chapter, suggests the adoption of a richer notion of logical consequence.

AGAINST THE ADEQUACY REQUIREMENT: AZZOUNI

Jody Azzouni[123] has objected along two lines. First of all he rejects the correctness of the adequacy requirement proposed by Shapiro and Ketland: a theory of truth ought not to prove generalizations like T-Teor$_{PA}$. Secondly, he rejects the extension of the induction schema to the new language L_T. According to Azzouni it is correct to consider essential to truth generalizations *concerning* truth (clauses like Gen: "for any sentence φ,ψ the conjunction of φ and ψ is true if and only if φ is true and ψ is true"), but it is not essential that a theory of truth proves generalizations *about* particular truths (like T-Teor$_{PA}$). It is not the job of a theory of truth to establish what is true and why. Azzouni says: "Call Physical truth the truths in some physical language L_P that follow from some physical theory A_P: surely it is not required of the deflationist that when she supplements L_P with a theory of truth that it should follow from her theory of truth that everything that follows from A_P is true[124]". The point is simply that the ability to prove non logical truths and generalizations goes rather beyond what

[123] Azzouni 1999.
[124] Azzouni 1999, p. 542.

a deflationist calls a deflationary theory of truth. It does not matter whether the generalizations in question are obvious or natural. In the case of mathematics, or PA, perhaps, a deflationist can accept that it is worth adding resources in order to characterize the arithmetical truths that are beyond a deflationary truth. However, once we permit this, what we get is not a deflationary theory anymore. Thus, it should not be surprising that we have lost conservativeness at the same time. At most, we should admit that in the case of mathematical truth we do not deal with deflationary truth. The idea of Azzouni seems the following: since we do not expect it to be a task of a theory of truth to prove something like: "everything the Pope says is true", we should not expect the truth theory to prove "everything PA proves is true" either. The only difference that could make us willing to prove the last statement is that it is more granted and desirable[125]. However, once we have accepted a Pope theory (which we can think as a schematic theory like: if the Pope says "φ", then φ), call it P, we want, at least, to be able to say that P is true, namely that everything the Pope says is true. Azzouni does not deny this point: a deflationary theory gives us enough resources to *say* this. What it must not do is give us resources to *prove* it. After all, "everything the Pope says is true" is very probably a false sentence and we do not want a theory, let alone a theory of truth, to prove anything false.

Although these remarks sound reasonable, they seem based on confusions. If our Pope theory P extended by a theory of truth T proved that everything the Pope says is true, would it be really a fault of the theory of truth? No. The error would lie in accepting the theory P. We should reject P. It is because we do not accept P that we reject

[125] See Shapiro 2002 for a Jewish version of the argument, in which what is true is everything the Rabbi says.

"everything the Pope says is true", not because our theory of truth is incorrect or too strong. We know that P is not a good theory. As we apply our truth theory T to a base theory B, T does not know if B is true or not, just because, as Azzouni pointed out, it is not a task of a truth theory to say which things are true. The theory of truth "trusts" us and the fact that we accept B. To construct a truth theory in such a way that once it is added to a theory B it is not able to prove that B is true is just an extreme move. The problem is the acceptance of B and not the extension of B with T: if we accept the former step, the latter seems justified. We can clarify the point putting it in a conditional form:

if we accept B, then B ∪ T ought to be able to prove the truth of B.

Why should we not adopt the adequacy requirement in such terms? In the case of a Pope theory, for instance, what would prevent us from getting "everything the Pope says is true" is that we have not a justification for the antecedent. The case of PA and mathematics is a different case because they are theories we are willing to accept and we do have justifications for them. We have no worry to state the antecedent and to derive that PA ∪ T ought to prove $T\text{-}Teor_{PA}$. What Azzouni says could be simply reduced, then, to the sceptical idea that we have no good reason to accept a base theory B.

Note that Azzouni's objection does not apply only to infinite generalizations. If it is not a task of a theory of truth to state particular truths, we should prevent it from proving, for instance, $T(\ulcorner\forall x\neg(Sx=0)\urcorner)$[126]. We could say that it is not a truth theory that should tell us that the first axiom of PA is true. The same ought to be said about finite generalizations that, instead, even the weakest deflationary theory (like

[126] Notice that $T(\ulcorner\forall x\neg(Sx=0)\urcorner)$ can be immediately deduced from the relevant biconditional.

DT|) can prove. It seems that in Azzouni's view a theory of truth should be unable to ascribe truth to anything.

What is really at stake in the adequacy requirement are the logical resources that a theory of truth should have, not what particular sentences belong to it. We expect, for example, a truth theory, when added to PA, to be able to prove T-Ax_{PA}, because we can prove the truth of each axiom of PA. We would like to have a theory of truth with enough strength to put these ascriptions together in a single generalization. Finally, why does Azzouni claim the legitimacy of generalizations such as "for any sentence φ,ψ the conjunction of φ and ψ is true if and only if φ is true and ψ is true"? To call these "generalizations *concerning* truth" does not change anything. We could see them as a particular truth: a truth about (classical) conjunction of sentences.

A second, more interesting, objection against the argument of Shapiro and Ketland concerns the induction schema and the opportunity of allowing the truth predicate in it. The argument from conservativeness underlines the tension between a Tarskian theory and the loss of conservativeness delivered by full induction. Shapiro cites Dummett's arguments in favour of full induction: it is part of the very concept of natural numbers that induction holds for every well defined property[127]. It follows that if we introduce a new predicate P with a determined extension on natural numbers, then we have reasons to let it enter the induction schema. The problem is that, Azzouni claims, such considerations hold only for the standard model $\mathbb{N}$: "It is simply not true of other models of A that any predicate P (however defined) with a determinate extension over the numbers of that models belongs in the induction scheme[128]". The existence of non standard models is a consequence of the

[127] Shapiro 1998, p. 500.
[128] Azzouni 1999, p. 543.

fact that, although the induction holds for every predicate definable in L_{PA}, this is not true of L_{PA} enriched with a truth predicate (in a tarskian style, at least). Some models of PA are not models of PA plus a (Tarskian) truth predicate and full induction (like T(PA)). A deflationist has just to remind that PA does not characterize the standard model despite others. Shapiro is right only if he has implicitly adopted the idea that some first order axiomatization can exactly capture $\mathbb{N}$. Only this way he can claim that the notion of truth is implicit in arithmetical concepts.

These remarks also help us make clear why Azzouni does not want a generalization like T-Teor$_{PA}$ to be derivable: because it is not true in every model of PA. We might want to prove T-Teor$_{PA}$ only if we have decided that PA characterizes the standard model. But we have no reason to do that. The same holds for the Gödelian sentence G. Is G true? Yes, but in the standard model $\mathbb{N}$. If $\mathbb{N}$ is not considered, we have no reason to consider G true. PA has different models, not isomorphic, and nothing in PA allows us to favour one of them. Since there are sentences (in L_{PA}) that are not true in every model of PA, a truth theory should not prove that they are true. If we accept PA we are committed to every model it characterizes, even non standard ones where G or T-Teor$_{PA}$ are false. If we accept all these models we might say that the sentences that PA makes true are at most those that are true in every model of PA: the theorems of PA. Since G is not always true, we should not consider it true nor should we prove it with a theory of truth. A truth theory should not make more sentences true than the base theory does. In this form, conservativeness seems to be just an adequacy requirement itself.

The idea that what is true in the standard model could be so ignored has been criticized by Murzi and Rossi[129]. They

[129] Murzi and Rossi 2020.

point out that sentences such as G are, as a matter of fact, commonly recognized as mathematical truths, so that also the deflationist should vindicate them. But if a deflationist accepts G on the grounds that it is a standard piece of mathematical knowledge, she should also accept other standard pieces of mathematical knowledge. This easily leads to accepting mathematical theories that re-introduce a non-conservative notion of arithmetical truth. For example, a deflationist should accept a theory like ACA[130]. Since ACA and CT are intertranslatable, by accepting ACA, also a strong non conservative theory of truth like CT is eventually endorsed. It then seems that deflationists cannot retain conservativeness, unless they reject standard pieces of mathematical knowledge at the same time. To Murzi and Rossi, however, it could be objected that a deflationist has some room to accept ACA without accepting CT as a theory of truth. Deflationists could insist that CT is not a theory of truth (because it is not purely disquotational), or that it is not *only* a theory of truth having also some mathematical content (following Field).

Dan Waxman[131] has defended and extended a strategy similar to Azzouni's. He argues that the conservative argument fails to show the inadequacy of deflationism. Take the sentence G. According to the adequacy requirement, a theory of truth should prove G when added to PA. However, as Waxman contends, this is correct only if G is indeed a consequence of PA in the first place. The critical point then is understanding the notion of logical consequence according to which G is or not entailed by PA. The issue eventually boils down to whether arithmetics is understood in axiomatic terms, as a mere a first order theory like PA, or

---

[130] More details on ACA are given in the section below where Tennant's view is discussed.

[131] Waxman 2017.

in terms of a categorical conception of the natural numbers. Both options, Waxman argues, are unproblematic to the deflationist. The first sense is the one discussed by Azzouni, in which we have PA and its many models. According to this option the standard model, in which G is true, is just one among many others in which G is not true. Waxman then agrees with Azzouni that if arithmetics is understood in the axiomatic way there is no reason to expect a theory of truth to prove G, since G is not entailed by arithmetic in the first place. The adequacy requirement is misplaced and deflationism can be conservative. No problem for deflationism is in this first option[132]. What about the other option, involving a categorical conception of arithmetics? A categorical characterisation of arithmetics can be typically achieved by endowing first order logic with additional resources, such as second order quantification, infinitary rules, and so on. Such logical resources commit to a semantic consequence relation according to which G is a semantic consequence of arithmetic. Since G is true in the standard model, it follows that the adequacy requirement is arguably correct. If a categorical characterization of the standard model is possessed, and this is what arithmetics is about, then G should be considered true after all. Accordingly, contrary to what happens in the axiomatic understanding of arithmetics, G should be proved by a theory of truth. However, since G is already semantically entailed by the base theory, according to the stronger logical consequence characterizing the standard model, any theory proving G will also be semantically conservative. Again, deflationism can be conservative without being inadequate. Notably, this second strategy was already considered by Shapiro and hastily dismissed by deflationists such as Halbach, who considered the move more a trap than a way out. The main reasons for

[132] Murzi and Rossi 2020 address Waxman explicitly too.

dismissing such a view are the dubious ontological profile of higher order logics, and the complexity and non-effective nature of the logical consequences relations involved[133]. Waxman replies to these worries arguing that, at bottom, they concern the very possibility of a categorical conception of arithmetic, rather than deflationism. Moreover, the strategy is not intended to show that deflationism should adopt a categorical characterization, but that, even if arithmetics is understood in a categorical way, deflationism is not in trouble. Indeed, no matter how arithmetics is understood, deflationism seems safe. If arithmetics is understood in a categorical way, then G is a semantic consequence of arithmetics, and conservativeness is retained. If, on the other hand, arithmetics is understood in an axiomatic way, then the adequacy requirement is wrong, as G and its cousins should not be proved, and conservativeness can be retained again.

The crucial point in such considerations is how we should interpret truth with respect to PA. Azzouni and Waxman consider what models make G true and evaluate a theory of truth accordingly. However, what we care about in the adequacy requirement is something else. At bottom, the adequacy requirement seems grounded in taking PA to be a *faithful theory*. It is because we trust PA that we want to prove $\text{T-Teor}_{\text{PA}}$.[134] Such two sides are not equivalent: to make explicit what is implicit in the acceptance of a theory is not a trivial business[135], and it is not reduced to a simple repetition of the former theory (as we are going to see in the next section). The points that need to be addressed, then, are: what does it mean to accept a theory like PA?

[133] See Hyttinen and Sandu 2004 for a discussion of higher logic consequence relations from the perspective of deflationism.

[134] Waxman might reply that this is correct, however, only if the categorical view has been favoured.

[135] See Nicolai and Piazza 2019.

What is the role of truth in manifesting its reliability?[136] A deflationist should show that truth has no role here or that it is compatible with conservativeness. This is what Neil Tennant tried to do.

AVOIDING THE ADEQUACY REQUIREMENT: TENNANT

Neil Tennant, in his article "Deflationism and Gödel phenomena"[137], focuses on the Gödelian sentence G and the phenomena connected to Gödel's theorems. Independently from the explicit use of the notion of conservativeness, Gödelian phenomena seem to be immediately relevant for deflationism. Deflationism has its root, Tennant says, in the Ramseyan idea according to which to assert that a sentence φ is true is equivalent to the assertion of φ. Apparently, then, there is no difference between a true and a justified assertion, or, in general between truth and proof[138]. Now the traditional philosophical interpretation of the first Gödel's theorem shows the gap between these two notions. Tennant cites Dummett: "by Gödel's theorem there exists, for any intuitively correct formal system for elementary arithmetic, a statement "G" expressible in the system but not provable in it, which not only is true but can be recognised by us to be true…[139]" In particular, the so-called semantic argument[140] for G reveals that our recognition of the truth of the Gödelian sentence involves the notion of truth, revealing its substantiality. Tennant wants to reconstruct this argument in a way that is acceptable to a deflationist and, at the same time, as close as possible to the original structure of

136 Murzi and Rossi 2020 also stress this point.
137 Tennant 2002.
138 Tennant 2002, p. 552.
139 Dummett 1963.
140 See infra Chapter three.

the argument. In this way, Tennant wants to show that a deflationist has sufficient means to vindicate the claim that G should be asserted and not denied.

Tennant cites[141] many versions of the semantic argument to find the kind of proof that a deflationist must reconstruct. The informal argument is finally sketched in the following way: "G is a universally quantified sentence[142] (…), every numerical instance of that predicate is provable in the system S[143]. (this claim requires a subargument exploiting Gödel numbering and the representability in S of recursive properties). Proof in S guarantees *truth*. Hence every numerical instance of G is *true*. So, since G is simply the universal quantification over those numerical instances, it must be *true*". This argument is the base of what Tennant pompously calls the substantialist dogma: the way in which the semantic argument establishes the truth of the Gödelian sentence requires the notion of truth to be substantial. Tennant's goal is to argue against this dogma by reformulating the semantic argument in a way available also to deflationists. It is worth anticipating that the proposed reconstruction shows that any resort to the notion of truth is avoidable and the argument is not semantic at all. This would strengthen the conclusion that the argument does not force a substantial notion of truth, since truth has no role at all.

To reconstruct the "semantic" argument in a rigorous way we have to enrich PA with the external tools that

---

[141] Tennant 2002, p. 555-556. Tennant cites Dummett, Kleene and Lucas. Quite curiously, he does not cite the informal reconstruction of the proof that Gödel himself sketched in the introduction to his theorems.

[142] We built G as equivalent to the sentence in L_{PA} $\neg\exists x Prov_{PA}(x,(Sub_{PA}(q, \ulcorner y \urcorner, q))$, which, however, is immediately equivalent, by interdefinability of quantifiers, to $\forall x \neg Prov_{PA}(x,(Sub_{PA}(q, \ulcorner y \urcorner, q))$.

[143] "S" is the name Tennant uses for a general formal theory of arithmetic.

are minimally requested to prove G and to respect the original structure[144]. Clearly, according to the first Gödel's theorem, such resources are not available in PA. Moreover, the required methods must have the exact strength, they must not prove more than is needed. They should prove exactly what it would be proved with the addition of G or Con_{PA}. Such considerations prevent us from following two strategies that are often adopted. The first is proving Con_{PA} and then G in a fragment of second order arithmetic, known as ACA[145] [146] (Arithmetic with Comprehension Axiom). ACA is a theory in the second order language yielded adding the following comprehension axiom to PA:

$$\exists X \forall y(y \in X \leftrightarrow \varphi)$$

where φ is a formula in the language of second order and in which X does not occur freely. First of all, the argument in ACA is carried out by defining a truth predicate for PA and then by following T(PA), so it makes use of the notion of truth. Not only, since ACA is known to be interdefinable with T(PA)[147] , the notions involved are clearly substantial. Another problem is that ACA is much stronger than what is requested for the derivation of G alone. Similar problems, Tennant notes, arise with a Σ^0_1-reflection for PA. This principle, which uses an explicit semantic vocabulary, says that every sentence Σ^0_1-provable in PA is true (in the standard model). It says that every existential quantification of a formula is provable in PA only if it has a witness among the standard natural numbers[148]. Also in this case both the

[144] This specification is important, because it would otherwise suffice to add G to PA.
[145] About ACA see Simpson 1998.
[146] Tennant does not mention ACA.
[147] In fact, T(PA) can define a notion of arithmetical comprehension. ACA and T(PA) are also known to have the same arithmetical strength, they prove the same sentences in L_{PA}.
[148] Tennant 2002, p. 563.

notion of truth and too strong notions are used. Another problem is that an argument relying on the Σ^0_1-reflection principle has not the same structure of the original informal argument. We would have the same problem also if we just added just Con_{PA}. Although Con_{PA} does not involve any semantic notion and it has the exact strength requested (Con_{PA} and G can be shown to be equivalent in PA), the argument would be too long and completely different from the informal one.

INTERMEZZO: REFLECTION PRINCIPLES

Discussing the strategy of Azzouni, we have enlightened that the desire to prove $\text{T-Teor}_{\text{PA}}$ comes from the desire of manifesting our trust in PA. In $\text{T-Teor}_{\text{PA}}$ an explicit mention of truth is involved. However, the resort to truth and $\text{T-Teor}_{\text{PA}}$ is not necessary to manifest our trust in PA. A first alternative way to express that PA is a trustworthy theory is that of claiming its consistency, adding the sentence Con_{PA} to PA. Con_{PA} can be proven even by stronger axioms, which partially express the reliability of PA using sentences in L_{PA}, with the form:

$$\text{RP: } \text{Prov}_{\text{PA}}(\ulcorner\varphi\urcorner) \rightarrow \varphi$$

for every φ in L_{PA}.

Such axioms are called reflection principles and they express the belief in everything PA proves. Clearly this makes our reliance in the system explicit: we think it is a good system and we trust everything it proves. Tennant cites the following remarks of Feferman: "A reflection principle provides that the axioms of the (extended system) shall express certain trust in the system of axioms (being extended)[149]", "By a reflection principle we understand a

[149] Feferman 1962.

description of a procedure for adding to any set of axioms S certain new axioms whose validity follows from the validity of the axioms of S and which formally express, within the language of S, evident consequences of the assumption that all the theorems of S are valid" and "Reflection principles are axiom schemata ... which express, insofar as is possible without use of the formal notion of truth, that whatever is provable in S is true[150]".

The addition of reflection principles to PA yields a non conservative extension of PA, since PA ∪ RP proves Con_{PA}, as it is easy to verify[151]. The addition of RP however is stronger than the simple addition of Con_{PA}. For example in the former but not in the latter case it is possible to prove: $Prov_{PA}(\lceil \neg Con_{PA} \rceil) \rightarrow \neg Con_{PA}$. Moreover, since Con_{PA} is provable in PA ∪ RP, in such an extension G is also provable. Feferman comments: "variant of Gödel's doctrine is that the 'true reason' for incompleteness phenomena is that though a formal system S may be informally recognized to be correct, we must adjoin formal expression of that recognition by means of a reflection principle in order to decide Gödel undecidable statements."[152]

It is also possible to strengthen RP imposing uniformity, as in the following version:

$$\text{URP: } \forall x(Prov_{PA}(\lceil \alpha(\mathbf{x}) \rceil) \rightarrow \alpha(x)).$$

Note that reflection principles, both in the local form RP and in the uniform form URP, are infinitely many. RP and URP are schemas that have infinite instances. If we want a universal closure, joining those instances together in a single axiom, we need the truth predicate to form the global reflection principle:

---

[150] Feferman 1991.

[151] Rigorously, the simple addition of such reflection principles yields a not conservative extension, since they are in L_{PA} but PA does not prove them.

[152] Feferman 1991, p. 233.

GRP: $\forall x(\mathrm{Prov}_{PA}(x) \rightarrow T(x))$

which is exactly T-Teor$_{PA}$. Clearly, if we introduce the truth predicate we have to add also axioms governing the behaviour of this new symbol. We can choose among different sets of axioms, the simpler ones in DT| and DT or those with a Tarskian inspiration like T(PA)| and T(PA). It is interesting to notice that if T(PA) is chosen, the addition of GRP is redundant, since T-Teor$_{PA}$ is already provable in T(PA). In the present context it would then be natural to consider T(PA) as theory expressing the reflective closure of a theory. Truth, as shaped in T(PA), is a device to manifest our trust in PA and in what it proves.

BACK TO TENNANT

Local reflection principles allow to express the correctness of a system like PA without employing the notion of truth. The idea of Tennant is that of resorting to such principles to reconstruct the semantic argument for G. URP, for example, is a good candidate since there is no semantic term occurring there. In the form proposed above, however, the formula $\alpha(x)$ can have any logical complexity, provided that it is in L_{PA}; URP then gives principles of arithmetical reflection. Tennant notices that this is more than we need and he suggests the following weakened version of URP:

URP_{pr}: $\forall x(\mathrm{Prov}_{PA}(\ulcorner \alpha(\mathbf{x}) \urcorner) \rightarrow \alpha(x))$

where $\alpha(x)$ is a primitive recursive formula. URP_{pr} has the exact strength needed: it proves nothing more than what the assumption of G alone proves.

The last step, and core of the project, is verifying that by using URP_{pr} it is possible to reconstruct the semantic argument in a rigorous way, respecting the original structure of the informal argument. So that: "it is not only the lightest

hammer to crack the walnut, but also the one that allows the user to swing his arm in the familiar way[153]". Now we are in the heart of Tennant's strategy, who, in an extension of PA, PA*, yielded adding URP_{pr} to PA, proposes a reconstruction of the (meta)proof of G, following the structure of the original informal argument. The (meta)proof assumes the consistency of PA.

PA*-proof of G:

suppose m codes a PA-proof, Δ, of G. By representability it follows that there is some PA-proof, Θ, of $Prov_{PA}(m,\ulcorner G\urcorner)$. Now Δ is a PA-proof of G, from which one can deduce $\forall x\neg Prov_{PA}(x,\ulcorner G\urcorner)$. By the elimination of the universal quantifier we have a PA-proof of $\neg Prov_{PA}(m,\ulcorner G\urcorner)$. So we get the contradiction $Prov_{PA}(m,\ulcorner G\urcorner) \wedge \neg Prov_{PA}(m,\ulcorner G\urcorner)$, against the assumption of the consistency of PA. Therefore, m does not code a PA-proof of G. If m does not code a PA-proof of G, then by representability it follows that there is some PA-proof of $\neg Prov_{PA}(m,\ulcorner G\urcorner)$. Since m here is arbitrary, for every n there is some PA-proof of $\neg Prov_{PA}(n, \ulcorner G\urcorner)$. (*)By URP_{pr}, it follows that – there is a PA*-proof Γ*of - $\forall y\neg Prov_{PA}(y,\ulcorner G\urcorner)$, that is equivalent to G. So there is in PA* a proof of G.

Tennant stresses that we could still suspect the truth predicate to be necessary in the omitted part, where it is proved that a PA-proof of G or a PA-refutation of G do not exist. This is not the case, however, and Tennant reproduces these (meta)proofs too[154]. Thus, the attempt at defeating deflationism, showing that it is unable to justify the sentence

[153] Tennant 2002, p. 573.

[154] Tennant 2002, p. 577-578. In particular, the proof that does not exist a refutation in PA of G is usually built using a Σ_1-reflection principle, which could seem indispensable. This, as Tennant shows, is not true, since URP_{pr}, which is implied by the Σ_1-reflection, is enough.

G and make sense of Gödelian phenomena, fails. Indeed, it is possible to make sense of these phenomena without using the notion of truth at all and making the proof completely available to a deflationist. What is usually explained through robust theories of truth can be explained, in a more satisfactory way, using more modest and less demanding tools.

A last point is worth noticing. Tennant[155] agrees with Shapiro that, in this way, a deflationist cannot say that every theorem of PA is true. In this sense a deflationist cannot state her will to assert any theorem of PA. What cannot be *said* by a deflationist, however, can be *shown*. What is explicitly said at the meta-level using the truth predicate ("every theorem of PA is true") can be shown adopting the corresponding inferential norm. The idea, with a Wittgensteinian inspiration, is that a deflationist can show, without any use of the truth predicate, what would otherwise need a truth predicate to be said. If a deflationist, however, were to say this explicitly (without proving it), she could still use a deflationary truth predicate or introduce a pro-sentential device (like: for every sentence S that PA proves, then *thatt*).

KETLAND'S REPLY TO TENNANT

Ketland has offered some interesting replies to Tennant. Ketland points to the exact formulation of the argument from conservativeness, showing that Tennant's attempt does not provide a way out. The problem, essentially, is that the entire argument put forward by Tennant misses the target. The original argument of Ketland and Shapiro was based, on the one hand, on the explanation of the insubstantiality in terms of conservativeness, and, on the other hand, on an adequacy

[155] Tennant 2002, p. 574.

requirement that a theory of truth ought to satisfy. Since the two points are incompatible, deflationism is doomed. The requirement of adequacy calls for a theory of truth that is able to *prove*, when added to a base theory, that everything this theory proves is true. In our case, where the base theory is PA, a truth theory is required to be able to prove $\text{T-Teor}_{\text{PA}}$. In other words, if we accept PA, and we have a previous grasp of the notion of truth, this must suffice to justify the claim that the axioms of PA are true, the rules of inference preserve truth, and therefore everything PA proves is true. All these claims should follow from the acceptance of PA and a good truth theory. If, on the contrary, a truth theory fails this requirement, it reveals its inadequacy. Notice that a derivation of such claims from a truth theory also gives a justification of them in terms of truth. $\text{T-Teor}_{\text{PA}}$ in fact is just the global reflection principle (GRP), from which, by T-sentences, it is easy to deduce weaker local reflection principles as URP, RP or URP_{pr}. This reveals the reflective nature of truth. The adequacy requirement is a consequence of this kind of remarks. It captures the idea that it is fundamental for a theory of truth to make sense of the reflective nature of truth, and such a nature is manifested, paradigmatically, in the ability to prove $\text{GRP/T-Teor}_{\text{PA}}$ .The reflective nature connects the notion of truth to what Ketland[156]calls the *conditional epistemic obligation.*

Conditional epistemic obligation:

"if one accepts a mathematical base theory S, then one is committed to accepting a number of *further statements* in the language of the base theory (and one of these is the Gödel sentence G)"

Ketland cites Feferman, who gives a nice explanation:

[156] Ketland 2005, p. 79.

"Gödel's theorems show the inadequacy of single formal systems… However at the same time they point to the possibility of systematically generating larger and larger systems whose acceptability is implicit in acceptance of the starting theory. The engines for that purpose are what have come to be called *reflection principles.*"[157] The connection with our problem is clear: truth has an essential reflective nature, so that an adequate truth theory should be able to prove the reflection principles. According to the conditional epistemic obligation, these are the means by which the commitments we take adopting a certain base theory are made explicit. Since the "further statements" the obligation speaks about are in the language of the base theory, any theory that makes them explicit makes lose conservativeness over the base theory. The argument from conservativeness, then, aims at showing that a deflationary theory of truth can not make sense of the reflective nature of truth[158].

Relevantly for Tennant's strategy, it does not matter whether we can give an alternative explanation, without any resort to truth, of the "further statements" G and Con_{PA}. What it should be done, instead, is showing how a deflationist can justify the reflection principles and the reflective nature of truth manifested in GRP/T-$Teor_{PA}$. A deflationist cannot just add these principles to deduce Con_{PA} and G because the point is justifying such principles, as the adequacy requirement demands. A simple assumption just pretends to solve the problem by ignoring it. Note that this option is not excluded. Both Shapiro and Ketland admit that: "Neither Shapiro nor I have denied that it might be possible for Tennant (or the deflationist) to provide some other non truth-theoretic[159]

[157] Feferman 1991, p. 1.

[158] See also Cieslinski 2010, and Tennant 2010 for a reply.

[159] Keep in mind that such an alternative explanation cannot be provided in truth theoretic terms because it would imply the non conservativeness of the truth theory, so it is not available to a

form of justification of reflection principles. But Tennant has not- at least not yet- provided any such alternative justification[160]." And: "I suppose that the deflationist can embed the arithmetic A in a richer mathematical theory, such as set theory (…) the claim would be that set theory provides the real explanation not arithmetic truth. Note however (…) that the deflationist has invoked a ton of new ontology just to avoid the notion of truth[161]". The risk here is that this move could be, again, more a trap than a way out: a deflationist would expel substantiality from truth just to put it into the new resources introduced. She would avoid a commitment to a substantial theory of truth by committing herself to some other strong theory like set theory.

Although Ketland's reply shows the inadequacy of Tennant's strategy, we can say on behalf of Tennant that the evaluation is a little unfair. First of all, the argument against deflationism had not been formulated in a completely clear way. For example, Shapiro, in the passage just cited, speaks about the sentence G and not about reflection principles: it is G, in Shapiro's words, that should be explained by deflationists not in terms of truth. Tennant does that with undeniable precision and success. Ketland himself spends a lot of words to explain the opportunity of a deduction of G from a Tarskian theory of truth (like T(PA)). Deflationism, Ketland claims, cannot make sense of this, therefore it can not explain every fact involving truth. What Tennant does, however, is showing, pace Ketland, how we can recognize the validity of G without any resort to truth. Indeed, his own reconstruction is arguably superior to Ketland's and Shapiro's because it respects the structure of the original semantic argument and it uses just resources of the exact strength needed.

---

deflationist.

[160] Ketland 2005, p. 87.

[161] Shapiro 1998, p. 506-507.

The issue of reflection principles, however, is more complicated. Reflection principles are the means by which our trust in a theory is made manifest, and from such principles we can deduce some commitments of such a trusting attitude. This attitude can be made completely explicit using a truth predicate, but it can be made explicit also by weaker reflection principles than GRP, like URP o RP. Tennant agrees that there is something that cannot be made here, since a deflationist cannot *prove* a principle like GRP/T-Teor$_{PA}$, and this is what Ketland objects to Tennant: such a proof is precisely what a truth theory should be able to give. Similar considerations probably hold for the Gödelian sentence too: we can see the truth of G. This is the core of the entire manoeuvre that Tennant does not satisfy. It is true that Tennant tries to explain G in an alternative framework but, even if we accept his reconstruction and justification of G (giving up the claim that an adequate truth theory should do the same), it remains that nothing similar holds for reflection principles. These are simply introduced without justification and this seems unacceptable to Ketland: "it is rather like saying that if we avoid explaining a phenomenon, we achieve 'philosophical modesty'. Probably the ideal way to achieve such 'modesty' in the scientific arena would be to abandon scientific explanation altogether[162]". Not only should reflection principles be justified, but arguably, for Ketland, should be justified in terms of truth. Tennant's attempt fails under both respects.

It can be wondered why Tennant misses the target in such a way. Probably, the great difference between their positions reveals, more than the immediate inadequacy of a strategy, the distance between their purposes. We have seen that Azzouni denies the adequacy requirement by attacking the requirement itself. G, in Azzouni's view, is simply not

[162] Ketland 2005, p. 85-86.

true, so that it should not be proved, let alone be proved by a theory of truth. We may want to obtain G only if we had already decided that we are dealing with the standard model. Azzouni however confuses between what is true in PA (what sentences are true in every model of PA), and what should be believed true, if PA is believed true (what sentences are true in every model of PA plus a theory of truth). That PA is true means that everything it proves is true. This leads us to accept Con_{PA} and G. Azzouni denies that. He thinks that PA and the claim that PA is true should have exactly the same models. How this can be kept together with the idea that our theory of truth is able to make sense of the statement that "PA is true" is, however, a mystery. Azzouni seems to think that it is possible to claim the truth of a theory without being committed to the claim that everything the theory proves is true. This can be kindly described as a clash of intuitions, although it seems more a *reductio ad absurdum*. The idea of Tennant is different. He does not deny that if we accept PA and we make our attitude manifest then we ought to manifest our acceptance of everything PA proves. What he denies is that the notion of truth is essentially involved here. We can manifest our attitude by reflection principles, as URP. Here is the bone of contention: why should we accept such reflection principles? How can we justify them? Ketland and Shapiro think this justification should be given in terms of truth, but Tennant disagrees. It is possible, though, that Tennant reasons in a similar way but from an opposite point of view: we introduce reflection principles to manifest our trust in PA. "Why should we accept reflection principles?", Ketland asks, "because we accept PA", Tennant could simply reply[163]. Our acceptance of PA has nothing to do with a truth theory; it has to do with numbers and arithmetic instead. The justification of reflection principles

[163] And he did reply that way in Tennant 2004.

is the same of PA: if we are justified to adopt PA then we are justified to consider PA a faithful theory and to manifest this attitude by suitable principles that need not involve truth. Indeed, Tennant could reverse the objection against Ketland requiring a justification for extending PA to T(PA). What is the justification for T(PA)? Any analysis must stop somewhere. It is not clear that stopping at T(PA) is better than stopping at reflection principle. The real point of disagreement then is not, as Ketland thinks, that Tennant does not justify the reflection principles. Rather, it is the fact that Tennant does not deduce these principles from a theory of truth. The core of the question lies in considering truth as a reflection principle. The alternative explanation of Tennant does not suffice to make a difference, since truth must be, on Ketland's view, a reflection principle. Any other justification of reflection principles, of Con_{PA} or G, cannot be satisfactory for Ketland and Shapiro because truth must be able to do that too. Here it is where Tennant disagrees, he thinks that the entire story could (and should) be reconstructed in another way. Truth, for Tennant is not reflective. Tennant certainly succeeded in showing how it is possible to keep many of the advantages of a tarskian theory of truth without any resort to the notion of truth. But is he also right to claim that truth is not reflective? Here deflationary and substantialist intuitions clash[164].

REFLECTIVE POWER OR CONSERVATIVENESS?

From a deflationist point of view it is natural to embrace conservativeness. And not only to clarify the insubstantiality of truth. One of the basis of deflationism is the idea that some form of equivalence holds between the simple assertion of a sentence p and the assertion that "p" is true.

[164] On reflective phenomena see Nicolai and Piazza 2019.

From this perspective, it is also natural to argue that such an equivalence holds also for sets of sentences or theories: to assert a certain theory should be equivalent to assert that such a theory is true. This equivalence leads directly to the conservativeness thesis: the ascription of truth to a theory B, being equivalent to B itself, should not prove new sentences in the language of B. On the opposite side, there are the considerations on the reflective nature of truth. As we have already stressed, to claim that a theory B is true forces the claim that everything B proves is true. Indeed, this statement is just a specification of what is meant by saying that B is true. How could it be possible to accept that B is true, but to deny that everything B proves is true or vice versa? The idea that truth behaves like a reflection principle is also natural. Apparently, *prima facie*, a deflationist has no reason to reject this. Nothing in the defence of the reflective power of truth seems to involve anti-deflationist ideas. However, dangers are in the neighbourhood. If we accept that a theory of truth, when applied to a base theory B, should prove that everything B proves is true, then we lose conservativeness. If the reflective power did not force a loss of conservativeness, a deflationist could make space to this aspect into her theory without worries. Since such two aspects are in conflict, however, deflationist/conservative intuitions and substantialist/reflective intuitions clash. Ketland argues that renouncing to the reflective features means renouncing to truth. Tennant, willing to defend deflationism, argues for the contrary: if we accept the reflective power we have to give up deflationist intuitions, so the former must be rejected. This result, however, is not on equal terms. It would be hard to consider deflationist and reflective intuitions on the same level. After all, deflationism is specific to a certain view of truth while reflective intuitions seem more basic and neutral. The best strategy for a deflationist, then, would be to vindicate both.

A first step in this direction can be made by developing the hypothesis sketched above, according to which it is natural for a deflationist to accept the equivalence between a theory B and the ascription of truth to B. Exactly as a deflationist claims that there is an equivalence between the assertion that p and the assertion that "p" is true. Although the passage might seem obvious, it is not innocent: in the latter case we deal with a single explicitly cited sentence, in the former we deal with theories, with sets of sentences. It is far from mandatory to hold that these two cases are of the same kind. It is entirely possible for a deflationist to insist that the equivalence between an explicit truth and the sentence itself does not force her to the equivalence between a theory and the assertion of the truth of that theory[165]. Indeed, modern deflationists have often argued in favour of the idea that the usefulness and the *raison d'être* of the truth predicate lies in enabling the expression of indirect endorsements and commitments on sentences we can not cite explicitly. The formulation of such expressions is not possible, or it is very difficult, without a truth predicate. This can motivate the thesis that the addition of a truth theory to a base theory allows to make explicit commitments that were only implicit in the base theory. The deflationist idea that truth works as a device to express indirect endorsements, and the extra-deflationist idea that truth has a reflective power share, under this light, the same basic intuitions. Obviously, if we can vanish the impression that a deflationist is committed to the equivalence between a theory and its truth, the problem of the loss of conservativeness (and hence the loss of insubstantiality) remains. So, although combining reflective power and conservativeness would be the perfect solution for a deflationist, so far everything tells

[165] Or between a blind ascription and the sentence such a blind ascription refers to. Chapter Eight, see infra, goes exactly in this direction.

us that this way is not viable. Hartry Field, however, thinks otherwise.

THE ROLE OF THE INDUCTION SCHEMA: HARTRY FIELD

Field claims that a notion of truth improves the expressive resources of a language[166]: "this is a point that deflationists (or those who call themselves that) like to stress. The main point of having the notion of truth, many deflationists say, is that it allows us to make fertile generalizations we could not otherwise make". Thus, when Shapiro argues that, for a deflationist, truth is metaphysically thin he should not mean that we cannot use it to make commitments on matters not involving truth. However, at the same time, deflationists, like Field, do not want a deflationary theory of truth to have not trivial consequences on subjects not involving truth. In other words, Field accepts that a deflationary theory must be conservative[167]. According to Field two aspects should be part of a deflationary theory: the ability to prove generalizations and conservativeness. Among the generalizations that Field has in mind, some are particularly important, for instance Gen.

As we know, it is entirely possible to add a truth predicate to PA in a conservative way. One of the simplest option is just adopting a theory like DT| or DT. These two theories, however, are not able to prove the desired infinite generalizations: they do not satisfy the first request of Field. Thus, conservative truth theories that include generalizations involving truths are more interesting: "since

[166] Remember, again, that this is an important difference with respect to redundantism.

[167] "there is no need to disagree with Shapiro when he says "conservativeness is essential to deflationism (497)" Field 1999, p. 536.

I think it is clear that without such general laws the truth predicate would not serve its main purpose[168]". In order to get this, T(PA)| should be preferred. T(PA)| proves exactly the generalizations Field desires, allowing conservativeness at the same time. A deflationist, Field says, can easily concede that not only T-sentences are essential to truth but also the generalizations involved in T(PA)|. This is confirmed by the fact that T(PA)| is still conservative.

Conservativeness is lost, however, in the moment we allow the truth predicate to enter into the induction schema, passing from T(PA)| to T(PA). Here then we have a problem: is not such a result a proof of the fact that, after all, the notion of truth, as axiomatized in T(PA)|, is substantial? To evaluate this subtle question, we have to consider what happens in the move from restricted induction (without the truth predicate) to full induction (with the truth predicate): "if the new induction axioms J involving truth are essential to truth, and logic is effectively codifiable[169], then the notion of truth is substantial (not deflationary)[170]". Certainly the axioms of full induction are important, since, according to the adequacy requirement, we should be able to prove that everything PA proves is true. Full induction is indispensable to this goal. Accordingly, we might think that the new induction axioms are essential to truth. This would drag us to the substantiality of truth, making the conservativeness of T(PA)| useless. Crucially, however, in Field's view, this is a mistake. To see why, what is meant by saying that the new axioms are necessary to truth must be clarified. The new axioms of induction are essential just in the sense that: 1. *they are needed if we are to arithmetically derive important facts that involve the notion of truth* and not

[168] Field 1999 p. 535.
[169] Namely if our logic is not a strong logic, such as second order or infinitary logic, as Shapiro 1998 proposed.
[170] Field 1999, p. 537.

in the sense that 2. *the truth of the new induction axioms depends only on the nature of truth*. Not only 2. does not follow from 1., but 2. is simply false. The induction schema, in fact, holds for every new predicate. New instances of the schema are obtained every time new symbols are added to the language, and not only if a truth predicate is added. The validity of the induction axioms is grounded in a property of natural numbers: that they are linearly ordered with each element having only finite predecessors[171]. The truth of the new induction axioms, then, depends only on the nature of natural numbers and not only on the nature of truth, as stated in 2. Therefore, the induction axioms can be considered essential to truth just in the sense of 1.. However, 1. does not allow the conclusion that truth is substantial. To such a conclusion 2. is needed. The idea is that truth would reveal its substantiality only if the axioms depending exclusively on the nature of truth (as those of T(PA)|) were not conservative. Since this is not the case, the responsibility for non conservativeness is on numbers, not on truth. To say it in a simple way: since it is just when we allow full induction that we lose conservativeness, the schema is guilty and the schema is grounded on numbers not on truth. That T(PA) is not conservative, then, does not matter for the substantiality of truth. Accordingly, there is no need not to reject T(PA) and full induction. A deflationist can accept the extension, passing from T(PA)| to T(PA), because the loss of conservativeness does not depend on truth and it is thus innocuous for deflationism.

The solution of Field certainly is the most attractive for a deflationist. If successful, it would satisfy all the proposed requirements and, at the same time, be able to explain the insubstantiality of truth in terms of conservativeness. T(PA), in fact, can prove T-Teor$_{PA}$, namely GRP, so that it can make

[171] Field 1999, p. 538.

sense of the reflective power of truth. On the other hand, the fact that T-Teor$_{PA}$ enables us to prove Con$_{PA}$ and G, is not a problem because the non conservativeness of T(PA) has been tamed and made harmless. The only conservativeness that matters is that of T(PA)|. A deflationist, then, can claim that her theory of truth, T(PA)|, is both adequate and conservative.

REPLY TO FIELD

Also Field's solution, in spite of his brilliant argument, suffers from serious troubles. The first problems concerns the axioms of T(PA)| and the dismissal of simpler T-sentences. T-sentences have an essential role in characterizing deflationism, and they are the basic source of deflationism. Although deflationism has many versions, and some proposals give up T-sentences, as the prosentential approach of Grover, it is always quite clear in which sense such alternatives can be considered deflationary. In the case of the Tarskian clauses that inspire T(PA)|, a deflationary reading is not so obvious. Note that we do not claim here that a Tarskian theory formalizes a robust notion of truth. We just want to remind that the status of this theory is controversial enough to motivate doubts about the viability of such axioms to a deflationist. Two simple facts confirm this: first of all, other authors, Davidson[172] for instance, have based their philosophical projects on axioms of this kind without drawing deflationary conclusions. Second, Field himself has initially characterized a deflationary theory using T-sentences and only because of technical problem he has embraced axiomatizations like T(PA)|. The moral is that until Field does not give us reasons to think that also T(PA)| is an authentic deflationary theory, it is possible that the

[172] Davidson 1984, 1990, 1996.

deflationist enterprise has been given up in this move.

The key passage of the strategy of Field, which discards the whole responsibility of the loss of conservativeness on numbers, is also not completely convincing. Consider again the fragment of second order arithmetic ACA[173] and its version with restricted induction ACA|. It is quite implausible to consider such theories insubstantial. Second order is commonly thought to bring heavy ontological commitments[174]. We could apply, however, the same argument of Field also to ACA|. Accordingly, it could be argue that ACA| axiomatizes unsubstantial notions since ACA| is a conservative extension of PA. This conservativeness is lost just in the moment we allow full induction, passing to ACA. But here we could reason as above: the real responsible for the loss of conservativeness is not ACA but the induction schema[175].

The problem we find with ACA helps us clarify what does not work in Field's strategy. Field is right to notice that the truth of the new axioms with full induction does not rely *only* on the nature of truth, because, clearly, the nature of numbers has an essential role. It is because (standard) natural numbers are that way that induction is true. What is wrong, however, is to completely overturn the assumption thinking that the truth of the new induction axioms depends *only* on the nature of (standard) natural numbers, as Field does. Of course, this is the claim Field needs to deny that the induction axioms are essential to truth, so that the loss of conservativeness in the move from T(PA)| to T(PA) can be put aside. This interpretation, however, is too strong. We can see this considering the move from DT| to DT. In this case, full induction has no effects on conservativeness; we cannot prove new arithmetical sentences in DT. We get,

[173] About ACA see above.
[174] But this point can be quite controversial.
[175] Shapiro 2002 makes the same observation.

however, new induction axioms. Indeed we get just the same induction axioms we obtain with T(PA). So why in the move from T(PA)| to T(PA), but not in the mode from DT| to DT, is conservativeness lost? Why is there such a difference? If Field was right, and the whole responsibility of the new axioms lay in the nature of natural numbers and not in that of truth, it would be legitimate to think that different axiomatizations of truth would make no difference with respect to natural numbers (that is with respect to sentences in L_{PA}). This is not the case: different theories have different consequences when combined with the same induction schema. In other words Field reasons: "when we pass from T(PA)| to T(PA), the only difference is full induction, thus it is full induction that is responsible for the loss of conservativeness. Full induction depends on numbers not on truth, therefore truth is innocent". Whereas we reply: "when we pass from DT to T(PA) the instances of full induction are exactly the same and the only difference is in the pure truth theoretic axioms, thus truth is the real culprit and numbers are innocent". The final interpretation is that both these arguments are quite right and the truth of the new axioms depends *both* on the nature of natural numbers and on the nature of truth. In this sense, unfortunately, the loss of conservativeness can be charged also on the theory of truth and it can no longer be discarded only on numbers. The strategy of Field then fails.

Volker Halbach has pointed out another objection to Field's argument. Field seems to admit that a deflationist should be committed to the conservativeness of the axioms that are essential to truth, but he has not such axioms. The only axioms he has are "mixed" axioms or purely arithmetical axioms (those of the base theory PA). The reason is that it can be shown that also the weakest theory of truth (DT|) has some arithmetical content, since it can prove that there are at least *two* objects[176]. The claim that *pure* axioms of truth

[176] See infra Chapter five.

are conservative, then, is trivially right, just because there are not such axioms[177].

GIVING UP CONSERVATIVENESS: VOLKER HALBACH

In the discussion of Field's strategy a particular contrast emerged: if a deflationary theory proves some infinite generalizations it can no longer be a conservative theory. This is important because truth generalizations are bound to the thesis that the truth predicate has its own *raison d'être* exactly in the ability to enable certain generalization. Volker Halbach[178], emphasizing this point, has argued against the conservativeness requirement. The core idea of his argument is rather simple: deflationists never embraced or mentioned conservativeness before it was proposed by the opponents, and since the theories that are adequate for deflationist purposes are non conservative theories, deflationism should just reject conservativeness. Although one of the main theses of deflationism is that truth lacks a robust nature, for Halbach this claim does not commit a deflationary theory to conservativeness. Indeed, before conservativeness was introduced, deflationists defended the idea truth is not a substantial property following a different strategy. The argument reconstructed by Halbach moves from the observation that a deflationist should accept that some uses of "is true" force to consider T-sentences necessary, a priori and/or analytic. The analytic nature of T-sentences gives rise to a problem about their translations. Consider the T-sentence ""snow is white" is true if and only if snow is white". It can not be translated, for instance in Italian, simply as ""la neve è bianca" è vero se e solo se la neve è bianca". Since

[177] Halbach 2001a, p. 88.
[178] Halbach 2001.

the expression "snow is white" is a name for the sentence between quotation marks, this sentence is just mentioned and it must not be translated. Hence, our translation must be ""snow is white" è vero se e solo se la neve è bianca". But this cannot be considered analytic for an Italian competent speaker. An explicit reference to the language does not help either, because ""snow is white" is true in English if and only if snow is white" is analytic, but the translation ""snow is white" è vero in Italiano se e solo se la neve è bianca" is not. Since correct translations should preserve the modal status of the sentences involved, the only option is to come back to the first translation ""la neve è bianca" è vero se solo se la neve è bianca". If so, the two truth predicates ("is true" and "è vero") have different extensions (one applies to English sentences, the other to Italian sentences), and they must ascribe different properties (if any). Indeed, if it is assumed that "is true" should be translated by "è vero", it follows that these two predicates cannot ascribe a property to sentences, because the properties expressed by "is true" and "è vero" are different. If they could be translated, then the property they express would be the same and they should have the same extension. Since they have different extensions, they cannot express a property of sentences. The point can be put in this Quinian way: since ""snow is white" is true if and only if snow is white" is necessarily and analytically equivalent to "snow is white", then the first sentence must be about snow too. Regardless of the validity of such an argument[179], Halbach reminds that it is on similar

[179] First of all, it makes a difference if we focus on necessity instead of analyticity (a notion that is not free from problems). Second, that "snow is white" is a name that only mentions the sentence and we should not translate it neglects the nature of quotation marks. The name in question is not an unanalysable expression. Indeed by quotation marks we can recover the sentence named and vice versa. Finally, we might equally argue that, since " "snow is white"is true" and "snow is white" are necessarily equivalent, both speak about the

grounds that deflationists have argued that truth is not a genuine predicate or an authentic property, and nothing in this approach involves conservativeness. Deflationism just claims that truth is a device for simple logico-grammatical purposes, like blind ascriptions and generalization. To be loyal to deflationism, we should focus on this.

Halbach thus proposes to definitely abandon conservativeness. A deflationist should admit that her theory can have substantial consequences, and that deflationism is not committed to any form of conservativeness. What ought to be claimed is that truth exists only to formulate generalizations and that a deflationist had better formulate his theory in a strong, possibly non conservative way way. "After all, he has never said that the function of expressing and proving generalizations is trivial and requires only a weak conservative theory. Truth does not serve any further purpose independent of expressing and proving generalizations; but a device of generalizations is a powerful tool, and that it does not serve any further purpose does not imply that it is blunt too. The deflationist's account of truth is not innocent, but that does not mean that it is wrong"[180]. As we know, a truth theory based only on T-sentences (like DT) is not able to *prove* infinite generalizations. Halbach[181], however, has shown how also such a theory can *express* infinite conjunctions. Still, a deflationist may have reasons not to be satisfied with DT. Its weakness, in fact, makes very difficult to make sense of the effective use of the truth predicate in the common logical and philosophical practice[182]. The formal work of logicians, for instance, has focused on much stronger axiomatizations than simple T-sentences. Stronger axioms are necessary in order to completely satisfy

sentence "snow is white".

[180] Halbach 2001a, p. 189.

[181] Halbach 1999b

[182] See for example Tarski 1956, and Gupta 1993.

the deflationist ambition of being able to serve as a tool for infinite generalizations. This does not mean that a theory of truth should be able to prove every infinite generalization. Indeed, the only strengthening that has seemed natural is the one obtained by adopting Tarskian clauses like those in T(PA)| and T(PA). This, despite conservativeness, seems the direction to go.

REPLY TO HALBACH

Halbach holds that a theory of truth deserves being called "deflationary" as long as it axiomatizes a notion of truth with the only purpose of serving logico-grammatical aims. The problem is that, even if Halbach is right that having substantial consequences does not make the theory wrong, however, it seems to make it not deflationary. He denies that the peculiarity of deflationary truth should be read in terms of conservativeness. After all, a deflationist can propose an alternative explanation, insisting, for example, on the importance of the logical role as the essential mark of deflationism. However, this reaction is problematic. The argument from conservativeness seems to bring to surface a relevant aspect of the insubstantiality of deflationary truth. If truth is capable of substantial consequence, then truth looks substantial after all. We can certainly think that conservativeness does not tell us everything about insubstantiality, but hardly can we deny that it seems to tell us something about it. Of course this could be resisted. The arguments of Shapiro and Ketland could be wrong and be rejected. But, if so, we should be told where the mistake is. Halbach did not do this. Instead, he points out that it is impossible for a conservative theory to meet some deflationist needs. Thus, what Halbach has shown is, at most, that the logical function of truth cannot be combined

with insubstantiality.

Note that insisting that the insubstantiality could be interpreted by the claim that truth only serves logico-grammatical purposes, would be unconvincing. If having a device to make generalizations provides a powerful tool, we cannot separate (for now at least) the strength of this tool from the innocence of truth. After all, it is the notion of truth that gives us such a powerful tool. Indeed we could argue in the opposite sense: deflationist truth makes a certain logical function possible. This function, in all his strength, reveals substantial consequences, thus revealing the substantial nature of truth. If such an outcome was accepted, we would have turned deflationism into a primitivist position. Finally, also in the case of Halbach's proposal we can question the suitability for a deflationist of a theory as T(PA) instead of simple T-sentences. Pairing this with the refusal of conservativeness, and possibly of insubstantiality, it can be argued that what is proposed is just a new theory that weakly resembles a deflationary theory of truth.

If this is the epilogue, we should just ought to accept the argument from conservativeness: an adequate truth theory cannot be a deflationary theory. Shapiro himself states: "on such a definition, presumably, the deflationist would gladly accept any semantic, logical, and even metaphysical features of the notion of truth that flow from its role in generalization. I have no problem with deflationism so defined[183]" The problem is just that insisting on speaking of "deflationism" in this sense would not make much sense anymore.

[183] Shapiro 2002, p. 116.

TACKING STOCK

The argument from conservativeness has been criticized by deflationists along different lines. The big differences among the attempted strategies show, once again, the variety of the positions that are usually covered under the label "deflationism". The differences also show how vague the commitments of deflationism can be. The argument from conservativeness forces deflationists to take a stand over several particular and crucial problems. In conclusion, two main kinds of attitudes emerged: on the one hand we have the line Azzouni/Tennant, which attempts at avoiding or denying the validity of the adequacy requirement, on the other hand, there is the line Field/Halbach, which searches for an acceptable way to satisfy the requirement. The first strategy embodies the most radical form of deflationism, by defending a theory with minimal commitments. However, by renouncing the reflective power of truth, such a strand forces a deflationary theory into an inadequate theory. The second strategy opens the door to stronger theories, but it risks losing the insubstantiality of truth. Such a situation seems to confirm that the argument raised by Shapiro and Ketland moves from an authentic problem: to formulate a theory both deflationary and adequate is a very hard task.

PART THREE

CHAPTER FIVE
T-SENTENCES Vs CONSERVATIVENESS
PART I - Logic

According to the conservativeness requirement the typical deflationist claim that truth lacks a substantial nature is to be read in the sense that a deflationary theory is a conservative theory. This has been made precise at the end of the third chapter specifying the conservativeness requirement:

Conservativeness requirement:

if T is a deflationary theory of truth in a language L_T, for every base theory B in a language L_B, and for every sentence φ in L_B,

if $T \cup B \models \varphi$ then $B \models \varphi$.

Where " $\models$ " stands for the first order logical consequence relation.

DT| AND THE EMPTY BASE THEORY

Conservativeness over logic would ensure the complete neutrality and innocence of the notion of truth. Volker Halbach[184], however, has shown with a simple argument

[184] Halbach 2001a.

that DT|[185] (and thus every theory of truth we have been considering - since DT| is a subtheory of DT, T(PA)| and T(PA)) is not conservative over the empty base theory, which is first order logic with identity. DT| hence does not satisfy the conservativeness requirement.

5.1 Proposition:

DT| is not conservative over the empty theory.

Proof:

If "∀x(x=x)" and "∀x(x≠ x)" are names, respectively of the sentences ∀x(x=x) and ∀x(x≠ x); we have in DT| the two T-sentences:

i. DT| ⊢ T("∀x(x=x)")↔∀x(x=x) and ii. DT| ⊢ T("∀x(x≠ x)")↔∀x(x≠ x),

in pure logic with identity, ∀x(x=x) can be proved and ∀x(x≠ x) can be refuted:

1.a DT| ⊢ ∀x(x=x) and 1.b DT| ⊢ ¬(∀x(x≠ x))

then, by modus ponens between i. and 1.a and modus tollens between ii. and 1.b we get, respectively:

2.a DT| ⊢ T("∀x(x=x)") and 2.b. DT| ⊢ ¬T("∀x(x≠ x)").

from which, by ∧-introduction:

3. DT| ⊢ (T("∀x(x=x)")) ∧ ¬ (T("∀x(x≠ x)"))

Now by the principle of indiscernibility of identicals we get:

4. DT| ⊢ "∀x(x=x)"≠"∀x(x≠ x)"

and by universal generalization:

β. DT| ⊢ ∃x∃y(x≠ y)

This result can be obtained, in a similar way, also if

[185] Here we use the label "DT|" in a different way from previous chapters. In the second chapter we defined a truth theory as including PA in it. Here we consider DT| (or in general a truth theory) as the pure truth theoretic part of the theory. If, and how, this makes real sense is the problem faced in this chapter.

we adopt rules instead of axioms for truth, substituting T-sentences with rules for the introduction and elimination of the truth predicate, like:

$$\text{T-intr: } \frac{\varphi}{T(\text{“}\varphi\text{”})} \qquad \text{T-elim: } \frac{T(\text{“}\varphi\text{”})}{\varphi}$$

$$\text{T }\neg\text{-intr: } \frac{\neg\varphi}{T(\text{“}\neg\varphi\text{”})} \qquad \text{T }\neg\text{-elim: } \frac{T(\text{“}\neg\varphi\text{”})}{\neg\,\varphi}$$

(where φ does not contain “T”).

Clearly, in the derivation of β not only axioms involving truth (the truth predicate) are used, but also axioms and rules for logic and identity. However, if we blamed these for the loss of conservativeness, we would be led to trivialize the entire issue. According to this interpretation any axiom or rule involving a new symbol would be conservative, since nothing could be proved without logical resources[186] , and we could always blame the rules of logic and identity.

It is worth reflecting on how serious Proposition 5.1 is. According to it deflationism is not just an inadequate theory, as Shapiro and Ketland claimed. Indeed, this is not a case where conservativeness prevents deflationism from doing something we expect from a good theory of truth. The problem, here, is that a deflationist proposal is simply impossible because a completely conservative truth theory, able to satisfy the conservativeness requirement, cannot be formulated at all. Proposition 5.1 shows that we can not keep T-sentences and conservativeness (namely insubstantiality) together. We have now obtained a reductio ad absurdum of deflationism.

[186] Halbach 2001a, p. 179.

ESCAPING LOGIC?

Facing such a radical result, we could think that the problem lies in the requirement of conservativeness. Perhaps the formulation above was too strong. So formulated, the criterion prevents us from giving a deflationary theory whatsoever, and from making sense of the original spirit of the argument from conservativeness. The basic idea of that argument is to show that some commitments of deflationism prevent the theory from being adequate with regard to what we ideally expect. What we have here, instead, is a more direct and drastic argument. This argument has nothing to do with the elegant and sophisticated initial argument. If we want to do justice to deflationism and to the argument, a different, weaker requirement must be proposed.

Volker Halbach, in "How Innocent is Deflationism?"[187]takes this route reflecting on the proof of β. The proof has two key steps. Initially T-sentences force us to admit the existence of something that satisfies the truth predicate and of something that does not satisfy it. The commitment to this existential generalization is already implicit the moment we ascribe truth to something. The moment we formulate axioms treating truth as a predicate. It is the very formulation of Tarskian biconditionals that forces us to such an ontological presupposition. At this point, we conclude that those objects must be different, because there is at least a predicate that an object satisfies but the other does not. In other words, T-sentences imply that at least one truth and one falsehood must exist, and that these are different things. If we suppose that this is not the case we should admit that the same object could be both true and false, and this would deprive the T-sentences of any sense[188]. The moral of the story is that the truth of β comes

[187] Halbach 2001a. makes such considerations more explicit.

[188] If we accept the possibility that some objects are both true and

from the assumption of axioms able to govern the behaviour of the truth predicate: axioms treating the expression as a predicate able to classify sentence types[189]. In order to make this possible we need to presuppose at least that sentence types exist and that they have some minimal features. We can agree with Halbach when he says: "it is not surprising that the T-sentences logically imply that there are at least two different objects; for the T-sentences have been motivated on the background of an ontology embracing abstract sentence types (or their codes)"[190]. β makes just clear the presuppositions implicit in the theory.

It is clear now why demanding a universal conservativeness would prevent us from formulating deflationism. The conservativeness requirement is really too strong. From such considerations Halbach concludes that we have reasons to weaken the initial request in favour of a more careful requirement: if truth is not substantial, it should not imply anything apart from the presuppositions needed to formulate it[191]. Accordingly, it is not the empty base the theory on which conservativeness must be required. We have to reformulate the requirement demanding conservativeness over a base theory that makes the ontological assumptions necessary to formulate a theory of truth explicit. We need to include, in the base, a theory of the objects to which the truth predicate is applied. This leads us to consider a formal theory of syntax as included in suitable base theory. As we have seen in the second chapter it is useful to take PA to be this theory. Our requirement, then, can be rephrased in the following way:

---

false, adopting dialetheism, the proof of the proposition 5.1 is blocked. In such a way we could have a completely innocent deflationism.

[189] Here we assume sentence types as truth bearers, but a similar result holds for other truth bearers as well.

[190] Halbach 2001a, p. 182.

[191] "The truth theory should not produce more than one has sunk into it". Halbach 2001a, p. 182.

New conservativeness requirement:

if T is a deflationary theory of truth in a language L_T, for every sentence φ in L_{PA},

if $T \cup PA \models \varphi$ then $PA \models \varphi$.

Where " $\models$ " stands for the first order logical consequence relation.

PROSENTENTIALISM AND CONSERVATIVENESS OVER LOGIC

We have often noted that "deflationism" is a title for a variety of different proposals. Conservativeness over logic is a specific case where the differences reveal to be nothing but irrelevant. The prosentential theory of truth, for instance, can appear to be a candidate to avoid the problematic result above. According to prosententialism "that is true" is a prosentence. Prosententialism takes "that is true" to be a whole expression, in which "true" only occurs as a syncategorematic term, without an independent meaning. Thus "is true" is not an authentic predicate. Such an approach seems able to block[192] both the passages of the proof of proposition 5.1. If "is true" is not a predicate[193], and it does not stand for a property, it is not legitimate to apply existential generalization and indiscernibility of identicals. Indeed, according to the prosententialist interpretation, the logical form of a biconditional like:

"snow is white" is true if and only if snow is white

is something like:

[192] Similar considerations hold for C.J.F. Williams, who thinks that there are no such things as truth bearers.

[193] Actually Grover eventually admits that "is true" can be considered an authentic syntactic predicate. She denies, however, that it is an authentic semantic predicate.

snow is white, that is true, if and only if snow is white

or a little bit more formally:

$(p \wedge \text{thatt}) \leftrightarrow p$

In this form we are not allowed to apply existential generalization and the proof of proposition 5.1. can hardly be recovered. In Brandom's version of prosententialism, - where "is true" is an authentic predicate working as a prosentence-forming-operator, however, the proof can be reconstructed.

Regardless of whether prosententialist views fare better with regard to conservativeness over logic, it is worth noticing that conservativeness would be out of place in any case. Conservativeness is useful to clarify what a deflationist means when she claims that truth is not a substantial property. For crude prosententialism, however, truth is not a property at all. If we simply deny to be in presence of a property whatsoever, the resort to conservativeness is superfluous.

SOME PROBLEMS

Although Halbach's considerations are natural and address an important target, there are some complications. It can be well conceded that β informs us of nothing more than the presuppositions necessary to formulate a truth theory. It can also be conceded that such a result is not surprising. However, this does not suffice to make truth innocent in the sense required by the conservativeness argument. In fact, one could argue for the contrary conclusion following the same lines: we could say that our truth theory is absolutely substantial because it reveals its own necessary ontological presuppositions. If we accept the conservativeness requirement, we are just facing a violation

of it. Certainly, what the T-sentences imply is not a big deal, and we can easily see where it comes from, but why should this be a good reason to turn a blind eye? Such a move would turn the request for conservativeness into a generic and vague request for an harmless content. Something like: a deflationary theory can be non conservative if it does not prove anything very problematic. Clearly this would amount to giving up the intuition of conservativeness and to fall back into vague and confused characterizations of insubstantiality. The other suggestion, according to which if something is unavoidable then we should concede it, is not convincing either. If we are persuaded that what makes the deflationary truth insubstantial is being conservative, that a violation of conservativeness is inevitable makes the problem even worse. If so, the only reason to change the requirement would be to avoid a simple confutation of deflationism[194].

In a few words, Halbach proposes a comparison to further defend the idea that conservativeness should be demanded over some richer base than logic. Halbach points out that even simple notation presupposes the existence of expressions and is not free from ontological commitments. Consider the clause: *if two sentences q and r of the base language are true then their conjunction* $q \wedge r$ *is true*[195]. Even the notation "$q \wedge r$" presupposes the existence of a conjunction of q and r. If we chose tokens as truth bearers, the principle above should be given up, because the conjunction of two tokens might not exist. Since a truth theory should not be considered more ontologically

[194] "In any case, I conclude from these results that deflationism would not be a tenable position if it were meant to imply that the truth theory is conservative over logic."(Halbach 2001, p. 181).

[195] Note that here we could avoid mentioning truth. Consider a different example: if q and r are two sentences of the language, $q \wedge r$ is equivalent to $\neg(\neg q \vee \neg r)$.

committing than simple notation, non conservativeness over logic should be ignored[196].

There is, however, an important difference: a truth theory makes its own ontological commitments explicit, whereas simple notation does not. Hence, truth is not exactly as innocent as notation. Moreover, if we admit that the ontological presuppositions of expressions are already implicit in the notation, then we can argue that by T-sentences we have not revealed the assumptions needed to formulate a truth theory. Rather, we have made explicit something that was implicit in the formulation of the base theory instead. A truth theory would be enabling us to show the ontological impact of simple notation.

ESCAPING LOGIC, SECOND ATTEMPT

Although Halbach's considerations are not completely satisfactory, he notes an important point. The critics of deflationism, Shapiro and Ketland, have accepted that the argument should be reconstructed demanding conservativeness over a theory of expressions, rather than over the empty theory. A reason not to take advantage of a strong argument, as the *reductio ad absurdum* that would follow from proposition 5.1, probably is that the intuitions of Halbach are someway right. So, we may wonder whether there is a way to make them preciser. The basic idea is that a truth theory makes sense only when joined with a theory of expressions. If we accept this idea, then we can represent neither a truth theory nor a deflationary theory if such a theory of expressions is not available. In this way, β is still a simple violation of our requirement, but we could avoid the problems. It is true that a violation would condemn

[196] Halbach 2001a, p.182. Halbach, then, assumes that notation is not ontologically problematic.

deflationism to death, but a deflationist could argue that if she is not given a base theory of expressions, what we are dealing with is not the truth theory she has in mind. Therefore deflationism is not victim of the lack of conservativeness over logic: no theory of expressions, no truth theory, no truth theory, no objection against deflationism. This idea can be developed further by appealing to the disquotationalist intuition. According to such an intuition everything there is to say about truth is exhausted by the disquotational feature of T-sentences. What Tarskian equivalences tell us is that truth is just the inverse operation of quotation. What we mean adopting the T-sentences as axioms is something like:

from a name "p" of a certain sentence p: to assert that "p" is true is nothing more than to assert p[197]. In other words, if the quotation operation gives us a name for every sentence, the truth predicate allows us to come back to the sentence, erasing the effect of quotation. Now, if truth is just disquotation we need to know what quotation is. In particular the following point is essential: the correlation between a name "p" and the sentence p must be addressed. To do justice to the importance of T-sentences we need to know that "p" is exactly the name of p. It is such a correlation that is missing in proposition 5.1. There we have the list of Tarskian equivalences and a name for each sentence, but nothing tells us that the name on the left is a name of a sentence, let alone the name of the sentence on the right. We know[198], at most, that "p" is a name but we do not know of what it is a name. If we lack such information, the role of the truth predicate is not what is expected and the deflationist idea is lost. If we have not such information what the biconditionals say is that a certain object in a certain domain

[197] Obviously such an idea should be clarified.

[198] We can recognize the syntactic category of names, but DT| cannot prove that such expressions are names.

has a certain property[199] named by "T", under the condition that a certain clause holds. For example one of the axioms is:

$$T \text{ “}\forall x(x = x)\text{”} \leftrightarrow \forall x(x = x)$$

We must not be confused here by the fact that on the left the sentence on the right occurs. We just chose this notation to mirror the right intuition. The formation rules, however, tell us only that if p is a sentence of our language, then "p" is a name of our language, but we do not know of what it is a name. What we have is just an extension of our language with an infinite number of individual constants, without restrictions on what they could denote. This is confirmed by the fact that although the number of sentences is infinite and so is the number of names, the domain has not been changed. Indeed, a domain with just two objects suffices to satisfy our truth theory. Thus, the axiom above could be rephrased simply as:

$$T(c) \leftrightarrow \forall x(x=x)$$

where "c" is an individual constant and where all we know about c is that it is an object of the domain. The theory of expressions we need should be able to determine, given a name of a sentence, the sentence it names and vice versa. It should be able, for instance, to prove that every sentence has its own name, so that if two sentences are different, also their names are[200]. Only if such a theory of expressions is available a truth theory can be sensibly formulated.

Now Halbach's intuition can be articulated in a better way. Proposition 5.1 is not able to condemn deflationism because what it involves is not a deflationary theory of truth. If a theory of expressions is not available, deflationism is not

[199] It is a property in the weak sense that "T" is a predicate with an extension.

[200] The same holds for other truth bearers, like propositions for instance.

available either. Accordingly, conservativeness should be required over the smallest theory on which a deflationary theory makes sense.

DT| AND THE THEORY OF EXPRESSIONS: ANOTHER PROBLEM

Unfortunately, also so formulated the conservativeness requirement can not be accepted. The problem is that, if without a theory of expressions a deflationary truth theory cannot be formulated, then the theory of syntax should be a *part* of our formalization of deflationism. If we stick with the idea that truth is just disquotation, then a formalization of quotation integrates our truth theory. The adequate formal version of deflationism is obtained from the *union* of T-sentences with a theory of syntax[201]. The syntax theory should only be part of the truth theory, not a part of the base.

A way out from this further difficulty can perhaps be found in some considerations put forward by Horwich. Horwich, in *Truth*[202], argues that a good theory of truth should be a theory of truth and nothing else. We should not expect a theory to explain every fact in every field. At most, we should expect a truth theory to be able to explain everything in a field when joined with a theory of that field. If so, we could think that we should keep a truth theory distinct from a theory of expressions. After all, these are different theories about different subjects. The fact that a deflationary theory is not able to explain anything unless it is not joined with some other theory could also be taken to

[201] Even Tarski (Tarski 1956) treats the theory of expressions as part of his (meta)theory of truth, confirming that this is a common practice in logic.
[202] Horwich 1998b.

be a confirmation of its supposed innocence. Unfortunately, even putting possible worries aside,[203] Horwich's suggestion does not solve our problem. Here the point is giving an explanation that a truth theory is not able to give alone. The point is that we have good reasons to think that if we do not integrate the theory, what is obtained is not what is expected. A comparison with logic is useful. Certainly logic is an independent theory, so that it could be separated from a truth theory. After all, only this way we would have a theory that is just a theory of truth and nothing else. Otherwise we would have a truth theory plus a logic theory. It is clear, however, that such an ambition is misplaced. A theory completely separated from logic is not possible; logic is a part of every other theory. This does not mean that logic cannot be considered as a base theory but only as a subtheory. We can take a theory B to be a subtheory of another theory B1 and at the same time we can require B1 to be conservative over B. This is exactly what we did with proposition 5.1. This means that we have to stick with the requirement of conservativeness over logic and the problems of proposition 5.1. Indeed, the situation is even worse, because our truth theory together with a theory of expressions is highly non conservative over logic.

5.2 Proposition:

DT| $\cup$ PA is non conservative over the empty theory B0.

(Sketch of the proof:

in DT| $\cup$ PA[204] we can clearly prove β again. But we can

[203] Substantiality seems only to be moved from truth into other notions. Moreover, probably we could adopt such a move in order to show the insubstantiality of any arbitrary notion.

[204] Remember that here we consider DT| (or in general a truth theory) as the pure truth theoretic part of the theory. Below we will speak

do more. For each number n, we can prove that there are at least n objects. This follows easily from the fact that every number has a successor and that 0 is not a successor of any number).

A LAST (PROBLEMATIC) OPTION

Only one possible option remains in order to escape the non conservativeness over logic. We could think that the consequences of a theory of expressions are not substantial because their nature is merely linguistic. We should not be worried about non conservativeness over logic because the content of what is proved is not worrying. Even if there were a violation of the conservativeness requirement, nothing really *about the world* would be proved. On the contrary, the result would only concern language and its properties. However, what regards only language cannot condemn deflationism.

The problem with this option is that it is not so simple to specify this idea. In order to state it clearly, we should be able to distinguish sharply between what concerns linguistic matters and what concerns the world. We should be able to distinguish between a linguistic theory and an extra-linguistic theory. This is not a trivial task, though. Take PA, is it a theory of expressions or is it an extra-linguistic theory? We are considering PA as our base theory, as a theory of expressions, but it is an arithmetical theory too. PA is both a theory of expressions and an arithmetical theory. This holds for every theory with enough resources. If we took ZFC as our base theory of expressions, should we forgive everything ZFC proves?

Let us close this topic by briefly touching on a related and intricate issue strictly related to such conundrums. We

---

again of a truth theory as including PA.

claimed that a theory of truth including a necessary theory of syntax (namely PA) should be conservative over a base theory. In the standard formulation of the conservativeness argument, such a base theory is just PA. However, since PA is able to provide a suitable theory of syntax, PA plays the role of the syntax and of the base theory at the same time. Given how entrenched a theory of truth is with a theory of syntax, one might wonder if, in general, we should not sharply separate the theory of syntax from the base theory. Since the adequacy requirement concerns exactly the role of truth in our meta-theoretic reasoning, the identification of numbers and expressions plays a critical role. There are clear reasons to feel uneasy with the usual situation[205]. As Halbach writes: "Identifying numbers and expressions is a notational simplification at best, but in informal metatheoretic discussion the theory of syntax and the theory of natural numbers should be kept separate: expressions are not numbers." (Halbach, 2011, p. 316). This issue forcefully emerges when one tries to add a theory of truth to some weak theory (like logic) not able to describe its own syntax at all. How is that possible to add a theory of truth to those theories without radically affecting them? It is clear that a forceful addition of a syntax theory, when possible, likely impacts on the underlying ontology, making conservativeness impossible, as the case of logic and proposition 5.2 shows.

All these inconveniences could apparently be circumvented if theories of truth were reformulated according to a new setting where syntax is carefully disentangled from the base theory. As one might expect, keeping a theory of truth with a syntax theory describing the language of the base theory separated from the base theory imposes quite

[205] Heck 2009 includes a long and clear discussion of such a topic.

substantial complications in the framework[206]. Here I skip those details and refer to Nicolai and Leigh (2013)[207] for a precise treatment. I just assume that a theory of truth with a theory of syntax disentangled from the base theory is added to a separate base theory. Note that, in this approach, PA can still serve as both the syntax and the base theory, as long as the two roles are kept sharply separated. For simplicity, however, we can imagine that we have a theory of truth formulated in a proper and distinguished syntax theory, like a concatenation theory[208], for the language of the base theory PA. What is mostly relevant for our purposes is the impact that such a reformulation has on the conservativeness debate. In particular, it can be wondered whether deflationists can gain convenient results with respect to conservativeness. Although the area is not widely explored, it seems that not much is to be expected. Two results are noteworthy. First, conservativeness is lost if the induction schema of the abs theory PA is extended with the language of the theory of truth with disentangled syntax. Under this respect, the situation parallels the passage from T(PA)| to T(PA). At the same time, however, such an extension seems to go against the intention to keep the two aspects disentangled, so that the result should not be surprising nor welcome. More interesting is what happens in the second following case. In the usual metatheoretic reasonings that a disentangled theory of truth intends to replicate, logicians do exploit the fact that syntax can be arithmetized. It is such an arithmetization that is exploited, for example, in Gödel's theorems and in the reasoning showing that the sentence

[206] Leigh and Nicolai 2013 offer a technically detailed and careful treatment. I refer to them and just sketch the gist of the strategy and the moral to be drawn from such an approach. See also Craig and Vaught 1958 for an early hint of this, which can be traced back to Tarski in any case.

[207] See also Heck 2009, and Halbach 2011, p. 316-321.

[208] Like in Grzegorczyk 2005.

G is true. Expressions can be coded by numbers after all. Here is then the second result. If bridge principles from PA to the disentangled theory of syntax are added in the form of coding axioms, replicating the arithmetization of syntax, then, unsurprisingly, conservativeness is also lost again. For some suitable base theories (like PA) the equivalence of syntactic claim with an object-theoretic equivalent becomes provable, making the disentanglement irrelevant.

The final moral is that, granted the merits of such an inquiry to enlighten the intricacies of our metatheoretical reasoning and to offer a more careful formulation of a truth theory, disentangling syntax from the base theory does not seem in general able to open further room for manoeuvre to deflationism. From now on, then, we go back to the usual entangled approach and put the disentanglement aside as an unnecessary complication in the present context.

CHAPTER SIX
T-SENTENCES Vs CONSERVATIVENESS
PART II – Peano Arithmetic

In the previous chapter we showed that the deflationist ambition to have a universal conservativeness over any base theory must be given up in favour of a weaker request of conservativeness over a theory of syntax. Although the motivations are not completely conclusive and they need a more satisfactory formulation, the idea of a complete innocence of deflationism, confirmed by conservativeness over logic, is a thesis no one argues for anymore. Keeping this in mind, we assume that it is possible to defend the claim that a deflationary theory should not have substantial consequences apart from those following from the assumption of a theory of syntax. What we are now going to analyse is whether in this sense T-sentences can be squared with conservativeness. At first sight, the solution seems straightforward: there are axiomatic theories that are conservative over PA: DT|, DT and T(PA)|. Thus that a deflationist can propose a conservative theory seems already established. Obviously, we do not mean to deny that such theories are conservative over PA. However, it can be shown that the reasons demanding conservativeness can be naturally and convincingly used to argue also in favour of stronger requests than conservativeness, imposing demands that cannot be satisfied by those theories (or variants of them). Moreover, such theories are severe simplifications of

adequate truth theories, and if more adequate theories are considered, conservativeness becomes hardly attainable.

The chapter does not mean to show that it is impossible for a deflationist to elaborate a theory meeting the conservativeness requirement. What we aim at showing is that the combination of conservativeness and deflationism is so hard to obtain that it should not be pursued light-heartedly.

FROM CONSERVATIVENESS TO EXPANDABILITY

As already noted, Shapiro focuses on a semantic understanding of conservativeness. If the addition of a truth theory T to a base theory B had consequences over the models of B, this would reveal the ability of T to have substantial consequences and it would disclose the robust nature of truth. The idea of *having consequences over the models of B* is made precise with the request of semantical conservativeness. We can wonder, however, whether there is not another way to understand this idea, which agrees with the spirit, if not with the letter, of Shapiro's argument. In some passages, in fact, Shapiro does not talk of conservativeness but of *expandability of models* instead. Shapiro says, for instance: "we can put the situation model theoretically. Let M be any model of A. Then the T-sentences determine an extension for the new predicate T, and with this extension M can be extended to a model M' of A' in the language L'. Thus any model for a theory without a truth predicate can be extended to a model with one. This is more grist for the deflationist mill that truth is metaphysically thin"[209]. Arguing in favour of second order logical consequence he then writes: "The result is general. Let Γ be any theory that can express its own syntax. Add a new predicate T to

[209] Shapiro 1998, p. 497.

the language and to Γ one of the common theories whose consequences are the T-sentences. Call the new theory Γ'. Then any model of Γ can be extended to a model of Γ'. (…) It follows that Γ' is a conservative extension of Γ"[210]. In these two passages what seems important for a deflationist is not just conservativeness. What matters is expandability[211]of every model of the base theory to a model of the base theory enriched with a truth theory. Conservativeness matters only as a consequence of expandability. In Shapiro's words what insubstantiality of truth commits to is expandability. Only because expandability implies conservativeness is a deflationist committed to conservativeness. Indeed, by considering the notion of expandability we can construct an argument which follows the very same lines of the argument based on the notion of conservativeness. Suppose that Karl accepts a theory B in a language L_B without a truth predicate. This means that Karl is willing to accept all models that make B true. Indeed, we can say that Karl uses the theory B just to talk about those models. Now suppose that Karl adds a truth theory T to his base theory, so B ⋃ T is yielded. If not every model of B was expandable to a model of B ⋃ T, the simple addition of truth could exclude some of the original models that B describes. In this way truth would have substantial and extra semantic consequences (beyond those involving the truth predicate) revealing a robust nature.

The choice of expandability can also be explicitly connected with the alleged metaphysics of the property of deflationary truth. First of all, also modern deflationists

[210] Shapiro 1998, p. 509.

[211] Here Shapiro uses the word "extension" loosely. Rigorously he should talk of *expandability*. Roughly, we have an *expansion* of a model when we add new symbols, while we obtain an *extension* if we add new elements to the domain. It is clear that when we add a truth predicate we do expand a model.

concede that the truth predicate is a predicate and, as such, it has an associated extension. Truth is then a property in at least that sense. It is at this point that deflationists and inflationists disagree, with the former claiming that such an extension corresponds to a thin property and the latter contending that it stands for a thick one[212]. How to clarify such a metaphysical disagreement in a formal context? Here is a way to proceed. If a truth theory is model-theoretically conservative over a base theory, then, every model of the base theory can be expanded to a model of the base theory plus the truth theory. To obtain a suitable extension we only need to operate at the level of the language, not at the level of the things the language is about (namely the elements in the domain and their properties/relations). Irrespective of what and how these things are, there is room for an insubstantial truth property. An extension for the truth predicate that leaves everything extra-semantical in the model unaltered can be found in any base model. This is the gist of expandability. Thus, if the truth theory is semantically conservative, the truth property that is theorized is not substantial in this sense: it is unable to shape the items instantiating it in any (new) way, over and beyond what the base theory already states. In other words, the truth property does not restrict the range of states of affairs that are possible from the point of view of the base theory.[213]

Remarkably, this approach can also be related to an independent metaphysical interpretation of the insubstantiality of truth. That we can find a way to gather the items in any domain means that we can find an extension in the model by carving it irrespective of its natural joints.

[212] In Strollo 2018, I stress the distinction between a property and a concept of truth. I then propose expandability to make sense of the insubstantiality of the property of truth, and relative interpretability to make sense of the simplicity of the concept of truth.

[213] See Strollo 2014a for an extended discussion.

This approach aligns with the philosophical idea that the insubstantiality of truth could be interpreted in terms of abundance, exploiting the metaphysical distinction between sparse (joints carving) and abundant (not necessarily joints carving) properties[214].

This version of the argument, based on expandability of models, is not only legitimate but it also helps clarify in which sense and why a deflationist should be committed to such notions. In any case, Shapiro does not argue in terms of expandability. A reason could be that Shapiro seems to think, wrongly, that the notions of (proof-theoretic) conservativeness and expandability are equivalent. A confirmation could be found at page 497 of Shapiro 1998. There he says that we can add a truth predicate to a first order theory respecting deductive conservativeness and a little below he notices that the question can be put in model-theoretic terms using expandability, thus suggesting that they are equivalent. Or perhaps, more charitably, it is likely that Shapiro is aware of the non equivalence of these notions and he simply prefers to spell out the argument in terms of conservativeness (rightly) evaluating that this makes no big difference to his own goals. A clue that also in Ketland's work the notion of expandability is relevant can be found in the fact that Ketland uses expandability of models in order to prove the conservativeness of some deflationary formal theories.

Cieslinski, however, explicitly criticizes the choice of expandability, that he just dubs "semantic conservativeness".[215]

[214] See Strollo 2014b for a more extended discussion of this connection. See Asay 2014, Edwards 2013 for the purely metaphysical side.

[215] Cieslinski 2015. Cieslinski also excludes proof-theoretic conservativity. He thinks that the main ways in which proof-theoretic conservativity could be justified (via claims of explanatory or justificatory roles) fail. He argues that conservativity does not guarantee the non existence of explanatory or justificatory truth theoretic proofs; neither does non-conservativity imply the existence

Here are his main reasons. First of all, he points out that no sound textual basis for imposing a conservativeness demand is easily found in the classical deflationist literature. This reason however, is quite weak. Conservativeness has been proposed as a way to make the idea of insubstantiality of deflationary truth more precise. Since such an idea has always been put forward in quite vague and suggestive terms, it is impossible to trace conservativeness back to some textual evidence. Indeed, this is true of any possible explication of the alleged insubstantiality. One could always retort that, given the vagueness of deflationist claims, there would be no textual evidence of any precise proposal. If so, no explanation of insubstantiality could be forthcoming. Cieslinski acknowledges that and concedes that one could propose semantic conservativeness as an explanation at least consistent with what the deflationists actually wrote.

A second objection against expandability has to do with the standard model. Arguably, when arithmetics is concerned, the model of PA we care about is the standard model. Certainly we do not want a theory, let alone a theory of truth, to exclude such a model. This consideration, however, does not lend support to semantic conservativeness. Quite the contrary. If what we care about is the standard model, then non standard and arithmetically wrong models need not be conserved. Thus, semantic conservativeness seems pointless, if not just a wrong demand. Moreover, if semantic conservativity is defended with the goal of keeping the standard model, then deflationists rely on a notion of arithmetical truth that seems out of reach for a deflationist. Truth in the intended model goes beyond deflationary truth, so that deflationists cannot afford it. There are, however, some immediate problems with this line of thought. To

---

of such proofs. Since I favour semantic conservativity, I put such issues aside.

see what, it is useful to have Waxman's strategy in mind. Arithmetic can be conceived in two ways, in a categorical way, as being about a certain specific intended model, or in an axiomatic way, in terms of a first order theory like PA. Cieslisnki considers the first alternative and seems to assume that a deflationist cannot afford it. But it is not clear why. Even because he concedes that deflationists are entitled to the usual model-theoretic and set theoretic notions. As discussed about Waxman's proposal, the intended model can be characterized using some higher order logic, without apparently invoking any robust notion of truth. What, if any, is incompatible with deflationary truth here is not clear, so that Cieslinski's argument is not conclusive.

Cieslinski finally considers a possible defense of semantic conservativeness based on an axiomatic view of arithmetics[216], according to which expandability of all models is desirable because no model of PA should be excluded. He considers two arguments for that move and rejects both. One is in terms of the idea that some models are correct, but we do not know which one exactly, so we should avoid any risk by keeping them all. I find this perspective hard to square with the assumption of an axiomatic understanding of arithmetics. Some categorical idea seems to have sneaked back. Since I agree with the exclusion of this option, I put it aside. Cieslinski eventually considers the idea that all models must be really treated on a par. This is exactly what one should think if an axiomatic view of arithmetics is endorsed. Cieslisnki however, sees this option as particularly costly. In his opinion, some sentences like $\mathrm{Con}_{\mathrm{PA}}$ are indeed true, and models that make them false are just wrong[217]. This, however, can hardly be made sense of, if the axiomatic view of arithmetics is accepted. If every model is on a par,

[216] Although he does not present things this way.

[217] This point is somehow similar to Murzi and Rossi 2020.

then Con_{PA} could indeed be false. The argument, if it does not just beg the question, then would seem to show that an axiomatic view is just wrong. So, at bottom, Cieslinski rejects expandability because he assumes that a categorical conception of arithmetics, in some form, is the correct one, and deflationists cannot afford it. However, neither of these claims is fully convincing or sufficiently supported.

If a reconstruction of the argument in terms of expandability is both legitimate and opportune, it can be shown that this reformulation has far from trivial consequences over the entire debate. We have noticed that expandability and (proof-theoretic) conservativeness are not equivalent: they do not imply each other. In fact, expandability implies conservativeness but not vice versa. A little more formally, let B be a base theory in a language L_B and T a new theory in the language L_T; if every model of B can expanded to a model of B $\cup$ T, this implies that B $\cup$ T is a conservative extension of B. However, the contrary does not hold: if B $\cup$ T is a conservative extension of B, this does not imply that every model of B can be expanded to a model of B $\cup$ T. We see an important example of this fact below. Expandability is a stronger request than simple (proof-theoretic) conservativeness. It can happen that a certain theory T is (proof-theoretically) conservative over another theory B, whereas expandability fails. If we focus on (proof-theoretic) conservativeness it is entirely possible to miss this possibility. This is a serious risk in the case of the debate over deflationism, because there are reasons to think that expandability is the relevant notion, and that conservativeness becomes important only because of it. Hence, the danger is that a theory apparently acceptable to deflationist does not really satisfy the requirement of expandability of models.

Requirement of expandability of models:

if T is a deflationary theory of truth in L_T, then, for every base theory B in L_B which includes a theory of syntax, it must be possible to expand every model M of B to a model M' of B ∪ T.

TECHNICAL INTERMEZZO: SATISFACTION CLASSES AND RECURSIVE SATURATION[218]

1. Non standard truths

In the second chapter we have introduced some techniques and operations - the arithmetization of syntax - which allow us to translate a discourse about the syntax of a given language into a discourse about numbers and properties of numbers. In this approach the sentences of a language are made to correspond to natural numbers, and, for a lot of aims, identified with them. What we do in the case of a language like L_{PA} is finding a biunivocal correspondence between the set of sentences in L_{PA} and the elements of the domain of $\mathbb{N}$, the standard model of arithmetic whose domain contains all and only standard natural numbers. One of the immediate consequences of Gödel's theorems is that PA has, beside the standard model $\mathbb{N}$, also different models, non isomorphic to $\mathbb{N}$, so that we can call them "non standard models"[219]. Let M be one of these non standard models, what happens if M instead of $\mathbb{N}$ is used as a base

[218] The literature on satisfaction classes and recursive saturation is highly technical. For general reference see Kaye 1991, Engström 2002, Kotlarski 1991 Kossak 1985. Personally, I have to thank Fredrik Engström for his patience to explain to me the quibbles of satisfaction classes. I certainly owe what I have understood (if any) to him and his long mails.

[219] For a good brief introduction to non standard models see Boolos, Burgess and Jeffrey 2007, chap 25.

for the arithmetization of syntax? What happens if we code the expressions of our language using not the elements in the domain of $\mathbb{N}$, but those in the domain of M? What happens if also non standard numbers are used? The first consequence is that we would get, beside standard sentences (the sentences coded by standard numbers) new mysterious non standard sentences, coded by non standard elements in M[220]. Such non standard sentences are those non standard elements that the model M "thinks" to be sentences (non standard numbers that code sentences in the sense of M). For example, we know that the syntactic property of being a sentence is representable in PA, so that in the standard case we have that the set of the sentences is given by all numbers x such that $\mathbb{N} \vDash \text{Sent}_{PA}(x)$. In a non standard case we have that the set of sentences is given by all x (standard and non standard) such that $M \vDash \text{Sent}_{PA}(x)$. It is not easy to give a clear idea of what these non standard sentences are. We propose just an example. Consider the sentence in L_{PA} $(\neg 0=0)$: this is an example of a standard sentence that $\mathbb{N}$ (and then M) recognizes to be a sentence, and that can be identified with its standard natural Gödel number. Similar cases are $(\neg 0=0)\wedge(\neg 0=0)$ and $(\neg 0=0)\wedge(\neg 0=0)\wedge(\neg 0=0)$, where the number of conjuncts is a standard natural number, (2 and 3). If the number of conjuncts is a non standard number, however, for instance $(\neg 0=0)\wedge(\neg 0=0)\wedge \ldots \wedge (\neg 0=0)$ (where the dots "…" stand for *a* repetitions of the sentence $(\neg 0=0)$, where *a* is a non standard number) what is obtained is not a standard sentence anymore. Rather, we have obtained a non standard sentence that M (if it contains *a*) can recognize it to be a sentence but $\mathbb{N}$ cannot.

Regarding these non standard sentences a natural question whether and how they are true. We know that a truth predicate, "T", such that $\mathbb{N} \vDash T\lceil\varphi\rceil \leftrightarrow \varphi$ for every

[220] The fundamental work is Robinson 1963.

sentence φ, is not definable in L_{PA}. The same holds for a truth predicate for non standard sentences and models, such that $M \vDash \Sigma(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$, for every sentence φ in the sense of M (standard and non standard). Such a predicate has to be added also in non standard cases. What should such a predicate tell us about non standard sentences? Here it is where we introduce the notion of *satisfaction class*.

A satisfaction class[221] S over a model M is a set of ordered pairs of the form $<\ulcorner\varphi\urcorner, \underline{a}>$ where $\ulcorner\varphi\urcorner$ belongs to the set of formulas[222] in the sense of M (standard and non standard) and $\underline{a}$ is a valuation for φ. Therefore $\underline{a}$ is a sequence of elements in M, corresponding to the free variables in φ. We can take a satisfaction class to be a subset of M.

6.1 Definition[223]:

If M is a model of PA, a subset S of M is a *satisfaction class* if and only if:

1. every x belonging to S is of the form $<\ulcorner\varphi\urcorner, \underline{a}>$, where φ belongs to Form(M) and $\underline{a}$ is a valuation for φ;

2. the class $\Phi(S) = \{\varphi$ belongs to Form(M) $\mid \exists\underline{a} < \ulcorner\varphi\urcorner, \underline{a} >$ belongs to S $\vee \forall\underline{a}$ ($\underline{a}$ is a valuation for φ) $\longrightarrow < \ulcorner\neg\varphi\urcorner, \underline{a} >, \}$is closed under immediate subformulas;

3. if $M \vDash \varphi$, $\underline{a}$ and $\ulcorner\varphi\urcorner$ is the Gödel number of φ, then $< \ulcorner\varphi\urcorner, \underline{a} >$ belongs to S;

4. if ¬φ belongs to Φ(S) and $\underline{a}$ is a valuation for φ, then $< \ulcorner\neg\varphi\urcorner, \underline{a} >$ belongs to S iff $< \ulcorner\varphi\urcorner, \underline{a} >$ does not belong to S.

---

[221] Although it is called "class" it is a set.

[222] We talk of formulas instead of sentences following the general literature on the topic.

[223] See Kossak 1985, and Krajewski 1976.

5. if $\varphi \vee \psi$ belongs to a $\Phi(S)$ and $\underline{a}$ is a valuation for $\varphi \vee \psi$, then $< \ulcorner \varphi \vee \psi \urcorner , \underline{a} >$ belongs to S iff $< \ulcorner \varphi \urcorner, \underline{a'} >$ belongs to S or $< \ulcorner \psi \urcorner, \underline{a''} >$ belongs to S; where $\underline{a'}$ e $\underline{a''}$ are suitable valuations for φ and ψ, respectively obtained from $\underline{a}$. (similarly for $\varphi \wedge \psi$).

6. if $\exists v_i \varphi$ belongs to $\Phi(S)$, $< \ulcorner \exists v_i \varphi \urcorner, \underline{a} >$ belongs to S iff $[(v_i$ is a free variable of φ and $\exists b < \ulcorner \varphi \urcorner, \underline{ab} >$ belongs to S) or (v_i is not a free variable of φ and $< \ulcorner \varphi \urcorner, \underline{a} >$ belongs to S)]; where $\underline{ab}$ is a valuation for φ obtained from $\underline{a}$ and b, obtained substituting b to the i-th element of $\underline{a}$ (similarly for $\forall v_i \varphi$).

It is apparent that the clauses defining a satisfaction class correspond to a tarskian definition of truth (or better of *satisfaction*), and that the reading is made easier keeping in mind the axioms of T(PA). Satisfaction classes can be classified further based on certain properties.

6.2 Definition:

A satisfaction class S on M is *full* if for every $\ulcorner \varphi \urcorner$ belonging to Form(M) and every valuation $\underline{a}$ for φ we have that $< \ulcorner \varphi \urcorner, \underline{a} >$ belongs to S or $< \ulcorner \neg \varphi \urcorner, \underline{a} >$ belongs to *S*.

6.3 Definition:

A satisfaction class S on M is *partial* if and only if there is α belonging to $M \setminus \mathbb{N}$ such that every time $M \vDash \mathrm{Form}_{PA}(\ulcorner \varphi \urcorner)$ and $\underline{a}$ belongs to M, if $(\ulcorner \varphi \urcorner) < \alpha$, then $< \ulcorner \varphi \urcorner, \underline{a} >$ belongs to S or $< \ulcorner \neg \varphi \urcorner, \underline{a} >$ belongs to S.

The idea is just that a satisfaction class is full if for every formula φ it contains φ or its negation and, if the satisfaction class is partial this is true only for the sentences coded by a (non standard) number smaller than α. Since standard sentences have a Gödelian that is a standard natural number

and every standard natural number is smaller than every non standard natural number, it follows that every satisfaction class (full or partial) behaves in the same way (they are full) with respect to standard sentences. It is important to notice also that a satisfaction class, even if it is a partial one, has to decide some non standard sentence; otherwise we have not a satisfaction class at all[224].

6.3 Definition:

A satisfaction class is *inductive* if and only if the expanded structure (M,S) satisfies all the induction axioms for every formula in the language $L_S = L_{PA} \cup \{S\}$ (Where the new symbol "S" is governed by axioms stating that S is a satisfaction class)

Combining these definitions further classifications can be obtained, distinguishing, among full and partial satisfaction classes, those satisfaction classes that are inductive and those that are not. Although a model can have many different satisfaction classes, not every model of PA can have one: non *recursively saturated models*, in fact, do not have any satisfaction class.

2. Recursive saturation

To better understand satisfaction classes, let us give some minimal information in order to explain the notion of recursive saturation[225].

[224] I owe this important remark to Fredrik Engström.

[225] It is possible to give the following definitions also in model-theoretic terms instead of speaking of theories.

6.4 Definition:

If B is a theory, a *type* over B is:

i. A set P($\underline{x}$) of formulas containing a finite number of free variables $\underline{x}$ ("$\underline{x}$" then stands for a sequence of variables).

ii. Such that B $\cup$ {$\varphi(\underline{c})$ | $\varphi(\underline{x})$ belongs to P($\underline{x}$) }is consistent. (Where "$\underline{c}$" stands for a multiple of - possibly new - individual constants).

6.5 Definition:

A type P(x) is *complete* if and only if T $\cup$ P(x) is a syntactical complete theory (that is for every $\varphi(\underline{x})$, T $\cup$ P($\underline{x}$) $\vdash \varphi(\underline{x})$ or T $\cup$ P($\underline{x}$) $\vdash \neg\varphi(\underline{x})$).

6.6 definition:

A type P($\underline{x}$) is *principal* if and only if there is a single formula $\psi(\underline{x})$ such that T $\vdash \forall x(\psi(\underline{x}) \rightarrow \varphi(\underline{x}))$, for every $\varphi(\underline{x})$ belonging to P($\underline{x}$).

6.7 Definition:

If M $\vDash$ B, a type P($\underline{x}$) is *realized* in M if and only if there is $\underline{a}$ belonging to M, such that M $\vDash \varphi(\underline{a})$ for every $\varphi(\underline{x})$ belonging to the type P($\underline{x}$). Otherwise M *omits* the type P($\underline{x}$).

For the completeness theorem, then, if P($\underline{x}$) is a type over a theory B, then B has a model that realizes P($\underline{x}$). Similarly, if P'($\underline{x}$), P''($\underline{x}$)... are types over the theory Th(M) of a model M (that is the set of all the sentences φ such that M $\vDash \varphi$), then there is an elementary extension M' of M that realizes every types P($\underline{x}$).

6.8 Definition

A type P(x̲) is *recursive* if the set {⌈φ(x̲)⌉ | ⌈φ(x̲)⌉ belongs to P(x̲)} is recursive. (That is if the set of codes of formulas in P(x̲) is recursive; notice that it is the set of formulas that is recursive, not the formulas, which can have whatever complexity).

6.9 Definition:

A model M is *recursively saturated* if and only if every recursive type over Th(M) is realized in M.

A recursively saturated model can be thought as a "big" and "homogeneous" model. A non recursively saturated model O is a model where at least one recursive type (a recursive set P(x̲) of formulas) is not realized in O. This means that there are formulas φ(x̲), belonging to P(x̲), for which elements a̲ in O such that O⊨ φ(a̲) are not available. This can happen, for example, when the model is not "homogeneous" or it is not "big" enough. To make it such, O should be expanded to O' adding new elements a̲ with the desired features.

The fundamental fact now is that if a non standard model admits a satisfaction class, then such a model must be big and homogeneous in this sense: it must be recursively saturated.

6.10 Lachlan's theorem[226]:

If M is a non standard model of PA with a full satisfaction class, then M is recursively saturated.

[226] Lachlan 1981.

It is possible to get a similar result also for partial satisfaction classes:

6.11 Theorem[227]**:**

If M is a non standard model of PA with a partial satisfaction class, then M is recursively saturated.

This can be summarized by saying that if M is non standard, and it has a satisfaction class (it does not matter whether full or partial, or whether it is inductive or not), then M must be recursively saturated. Recursive saturation is a necessary condition for a non standard model to have a satisfaction class[228]. However, recursive saturation is not a sufficient condition to guarantee the possibility of a satisfaction class: in fact there exist non countable recursively saturated models without a full satisfaction class or an inductive satisfaction class[229]. Recursive saturation of a non standard model is a sufficient condition to have a satisfaction class only together with countability.

6.12 Theorem

If M is a countable recursively saturated model of PA, then M admits a satisfaction class.

Notice that this does not mean that every countable recursively saturated model of PA admits whatever satisfaction class. For instance this is not enough to have

---

[227] See Kaye 1991, Theorem 15.5 and proposition 15.4.

[228] Notice that this is not true for the standard model $\mathbb{N}$. $\mathbb{N}$ is not recursively saturated but it does admit a "satisfaction class".

[229] See Kaufmann 1977.

a full inductive satisfaction class. It is important for us to notice that there are non standard models that are non recursively saturated. Therefore, crucially for our purposes, there are non standard models of PA such that they do not admit a satisfaction class.

SATISFACTION CLASSES AND AXIOMATIC TRUTH THEORIES

With such results available we can draw some important conclusions. The first observation is rather natural and concerns the relation between satisfaction classes and axiomatic theories of truth. The notion of satisfaction class has been constructed with the purpose of characterizing the set of all truths in a certain model from a model-theoretic point of view, while the axiomatic approach tries to characterize the behaviour of the truth predicate. It is clear that such approaches can be considered, in a certain measure, as two perspectives on the same problem. We can then expect an axiomatic theory of truth to often give an axiomatization of a predicate that defines a satisfaction class and vice versa. Indeed, an axiomatization for the truth predicate can be obtained by turning the clauses of a satisfaction class into axioms. In this way, what is get is just the axiomatic Tarskian theory T(PA)| - if we do not allow full induction- or T(PA) - if we allow full induction[230]. T(PA) defines a full inductive satisfaction class and T(PA)| a full not inductive satisfaction class.

The cases of DT| and DT are more complicated. Both

[230] In our definition of satisfaction class we used a relation symbol to speak about the satisfaction of a formula by a sequence of objects, while in the axiomatic theories we are using a one-place truth predicate. This difference, however, is not relevant here. It would have been possible, for example, to define a satisfaction class avoiding the notion of satisfaction (as in Engström 2002).

theories do not define a satisfaction class, but we can get a satisfaction class from DT with a little variation. In order to define a satisfaction class a uniformity requirement needs to be introduced. Accordingly, DT must be turned into:

UDT: $\forall x(T(\lceil \varphi(\mathbf{x}) \rceil) \leftrightarrow \varphi(x))$

where $\varphi(x)$ is a formula in L_T that does not contain "T". UDT defines a partial inductive satisfaction class. Notice that every model of PA can be expanded to a model of DT| or UDT|. These theories in fact do not decide any non standard sentence[231], so that they do not define a satisfaction class[232]. Therefore DT| and UDT| satisfy the requirement of expandability of models[233].

By contrast, since all axiomatizations T(PA), T(PA)| and UDT define satisfaction classes, it follows that they can only have recursively saturated models. Let Γ be one of the theories among T(PA), T(PA)| or UDT, since Γ defines a satisfaction class then, if M is a non standard model of PA and expandable to a model of PA ⋃ Γ, M must be recursively saturated. This means that not every non standard model of PA can be expanded to a model of PA ⋃ Γ, but only (at most) the recursively saturated ones[234]. A side effect is that to prove the conservativeness of T(PA)| we cannot try to show that every model of PA can be expanded to a model of PA ⋃ T(PA)|, because this is not the case[235].

This situation has great relevance for deflationism and for the problem of the insubstantiality of truth. As emphasized, insubstantiality can be interpreted as expandability

---

231 See Definition 6.1.

232 To see this consider Th(M) - the set of (standard) sentences that are true in M - and define the extension of "T" (both for DT| and UDT|) in M as the set of (codes of) sentences that are in Th(M).

233 They are the only deflationary theories satisfying such a requirement.

234 Recursive saturation is not enough to allow a non standard model of PA to be expanded, for instance, to a model of T(PA).

235 Halbach 1999a.

of models. But then, since no theory Γ can satisfy this requirement, no theory Γ is available to a deflationist. That T(PA) is not available is not really surprising, since this theory is not conservative. But that theories like T(PA)| or even UDT[236] are not acceptable either is an unexpected result. In particular, even if the attempt of Field to use the conservativeness of T(PA)| succeeded, the strategy would be rejected by the fact that such a theory cannot meet the expandability requirement.

A possible reaction for a deflationist could be to stick to DT| or DT, rejecting both T(PA)| and UDT (the uniform version of DT). After all, a deflationist might remark, no one has ever adopted a deflationary theory like UDT, but only theories with local T-sentences,which do not define a satisfaction class. However, what reasons could a deflationist have to prevent the uniformity of T-sentences? The axioms of UDT have the same disquotational feature of DT. The only difference lies in the fact that such axioms are made uniform by a quantification that objectively quantifies into the sentences to which truth is ascribed. It is hard to maintain that this operation is problematic or in conflict with some deflationist thesis. Certainly UDT has not been proposed yet, but what does bother a deflationist in this little variation of T-sentences? The only real reason seems to be that a deflationist cannot adopt UDT because, although it is a conservative theory, it does not respect the expandability requirement. This however, is clearly *ad hoc*. Moreover that a deflationist must renounce to uniform versions of theories based on T-sentences (slightly different and innocent versions of DT) invites the to question on whether the T-sentences are really as innocent as they are claimed to be[237].

---

[236] Notice that UDT is conservative over PA. See Halbach 1999a.

[237] It is worth noting that another theory is available. Since there is an axiomatization that resembles the strength of T(PA)| but such that

DT IS NOT INNOCENT EITHER

It has been told that in order to obtain a satisfaction class from DT we need to ask for uniformity, passing from DT to UDT. Although DT does not define a satisfaction class, however, not every non standard model of PA can be expanded to a model of DT. As in the case of T(PA)|, this is another case in which it is clear that conservativeness does not imply expandability: DT is conservative over PA but not every non standard model of PA can be expanded to a model of DT. This is, however, an even more interesting case, since it does not involve satisfaction classes and recursive saturation.

6.13 Proposition[238]:

A non standard model M of PA can be expanded to a model M' of DT if and only if Th(M) is in SSy(M).

Where Th(M) = $\{\varphi \mid M \vDash \varphi\}$, for every (standard) sentence φ, and SSy(M) is the standard system according to M; that is the set of all subsets of the standard model that are coded in M. In other words, SSy(M) is the set of any subset $\cup$ such that there exists a in M that codes $\cup$. Th(M) (which is a subset of the set of codes of standard sentences so it is a subset of) hence must be coded in M. Note that not every non-standard model M of PA is such that Th(M) is in SSy(M), in fact Th(M) is clearly non recursive, but for each non recursive set S

---

does not define a satisfaction class, so that every non standard model of PA can be expanded to a model of it. Such a theory is called PT| (see Halbach 1999a, p. 357). Its axioms, however, are more elaborated, and it is not clear whether they can be taken to be really deflationary. Clearly, since the version of PT| with full induction, PT, includes DT it cannot satisfy the requirement of expandability.

238 See Strollo 2014a.

of standard numbers there is a non-standard model M in which S is not coded (Kaye 1991, 142, Lemma 11.2.).[239].

Proof[240]:

1. Left to right.

Suppose there is an expansion (M,Γ) of M that satisfies TB. Where Γ is the set giving the extension of the predicate 'Tr'. For every standard natural number n there is an a in M coding the set Γ ∩ {0,1,2, . . . ,n}, since such set is finite. Thus, by overspill, there must be an a′ in M

coding the set Γ ∩{0,1,2, . . . ,b} for non-standard b. Since b is non-standard the set {0,1,2,. . . ,b} is infinite and includes every standard natural number. Therefore we have that Γ ∩ {0,1,2, . . . ,b} belongs to SSy(M). But Γ ∩ {0,1,2, . . . ,b} is just Th(M), thus a′ codes Th(M) and Th(M) is in SSy(M).

2. Right to left

Suppose that Th(M) is in SSy(M) and a is the code in M of Th(M). Then we can interpret 'Tr(x)' by 'x is in (the set coded by) a′ .

Putting all of this together, it follows that every theory including DT or defining a satisfaction class cannot satisfy the requirement of expandability of models. Such theories, then, are not available to deflationism. We must exclude T(PA) (which includes DT and also defines a full inductive satisfaction class), T(PA)| (which defines a full non inductive satisfaction class) - pace Field - , UDT (which defines a

[239] In contrast with saturated models, *prime models* are models as simple as possible. While a saturated model realizes as many types as possible, a prime model realizes as few as possible: it realizes only the types which cannot be omitted and omits the others. See Kaye 1991.

[240] The proof, published in Strollo 2014a, is due to Fredrik Engström. Cieslinski independently proved the same result.

partial not inductive satisfaction class and includes DT) and DT. The only available candidates left are DT| (or its variant UDT|)[241].

The fact that a deflationist must renounce to such a great number of theories, which apparently are good candidates (and theories based on simple T-sentences like DT or UDT among them), is a very good reason to question whether the marriage between deflationism and conservativeness/expandability is possible. Moreover, DT| and UDT|, the only remaining candidates, are the weakest theories proposed so far and their adoption would make it very easy to argue for the inadequacy of deflationism.

Unfortunately, the problems are not even finished yet.

DISASTROUS T-SENTENCES

All theories we have been considering so far, and on which the debate has focused, narrow the applicability of the truth predicate to sentences that do not contain the truth predicate. This is the case of theories constructed over T-sentences (DT| and DT), or those with a Tarskian inspiration (T(PA)| and T(PA)). Such theories allow to ascribe truth only in the form T($\ulcorner\varphi\urcorner$), where φ is a sentence in which "T" does not occur. This restriction enables to avoid a great number of technical complications and to evaluate the proposal in a simpler way but, at the same time, it prevents these theories from giving a satisfactory account of the truth predicate as effectively used in natural languages and in logical and philosophical practice. A theory that does not allow to iterate the truth predicate is a theory that cannot be considered adequate or definitive. It is clear that sentences

[241] Other interesting results about expandability for more complex theories can be found in Cieslinski, Wcisło, and Łełyk (2017), Cieslinski (2017), and Łełyk, M. and Wcisło, B. 2017, 2019.

like ““snow is white” is true” is true” or ““snow is green” is true” is not true”, should find space in a good theory of truth.

Deflationism insists on the fact that having a truth predicate gives us, by T-sentences, an essential device for expressing commitments beyond those possible in a language without such a predicate. Take the sentence: “something the Pope says is true”, and suppose that the Pope asserts just one sentence: ““snow is white” is true”. If the deflationist does not give us the corresponding T-sentence - ““snow is white” is true” is true if and only if “snow is white” is true - the deflationist machinery to express a commitment to this claim cannot be provided. A T-sentence allowing us to quantify objectually over ““snow is white” is true” would be lacking. The deflationist explanation, in fact, is intended to show that the truth predicate enables us to quantify objectually over the infinite disjunction:

P: (the Pope says “grass is green” and grass is green) or (the Pope says “sky is blue” and sky is blue) or (the Pope says ““snow is white” is true” and “snow is white” is true) or …

Where the dots mean that the disjunction goes on with a disjunct for every sentence of the language. Thanks to T-sentences we are then supposed to be able to pass to the new infinite disjunction:

P’: (the Pope says “grass is green” and “grass is green” is true) or (the Pope says “sky is blue” and the “sky is blue” is true) or (the Pope says “ “snow is white” is true” and ““snow is white” is true” is true) or …

so that we can get the formula:

P’’: (the Pope says x and x is true) or (the Pope says y and y is true) or (the Pope says z and z is true) or …

where we can objectually quantify:

P’’’: ∃x(the Pope says x and x is true).

The problem here is that we can not pass from ““snow is

white" is true" to ""snow is white" is true" is true", if we do not have the corresponding T-sentence. Since this argument can be proposed for every arbitrary sentence, a deflationary theory must cover the possibility of every iteration of the truth predicate. After all, in our case, there is no reason to forbid the Pope to say ""snow is white" is true"[242].

Notice that this fact leads also to worse consequences than the simple inadequacy of the theory. Suppose the Pope said exactly two sentences: "grass is red" and ""snow is white" is true". The first sentence is false and the second is true, hence our existential generalization "∃x(the Pope says x and x is true)" should be true. However, since our simplified theory can cover with its T-sentences only the first sentence, where the truth predicate is not iterated, according to our theory only "grass is red" is considered. This sentence is false so that also our generalization is. The most natural (and probably the only really satisfactory) solution to this problem is that of allowing truth to be predicable of any sentence, even of sentences containing the truth predicate, without any restriction whatsoever. This means that we should modify our theory DT| into DT|*, whose axioms (beside those of PA) will be all sentences in L_T of the form:

$T(\ulcorner\varphi\urcorner)\leftrightarrow\varphi$

where φ is a sentence in L_T without any restriction.

A deflationist has also another simpler reason to accept DT|*. According to one of the most basic deflationist theses, T-sentences are the fundamental axioms for truth. They suffice to explain any fact involving truth and they do not need any further explanation. But if a lot of T-sentences are problematic and we should put them aside, why should we accept T-sentences instead of other more complicated formulations? It seems that there is something

[242] On similar problems see Armour-Garb 2004.

more, and crucial to be said about truth, if the T-sentences must be restricted in some ways. A view accepting all T-sentences is often introduced as the only theory of truth wholly natural and acceptable to common sense and pre-theoretical intuitions. It is often called the "naïve theory of truth". It is on the ground of such trivial features that deflationists argue in favour of the innocence of their own theory[243]. Unfortunately, because of the liar paradox, DT|* is inconsistent.[244] DT|* is not then a viable candidate for a deflationary theory of truth. Apart from the obvious inconsistency, however, here we want to stress a fact that, although trivial, is never remarked. Since DT|* is inconsistent, DT|* is not conservative over any base theory (apart from those base theories that are already inconsistent), because it proves any sentence. This means that the most natural and complete set of T-sentences is actually at the opposite of conservativeness. This is important because the conflict between T-sentences and conservativeness could not be clearer.

APPEARANCES CAN BE DECEPTIVE

DT|* is inconsistent for a well known reason: the liar paradox. So DT|* must be somehow limited. The liar paradox arises from a sentence that says of itself that it is *not* true. Hence one of the roots of the paradox is, in general, the interaction between negation and the truth predicate. Certainly many, if not most, of those interactions are not problematic and a good theory should allow them. Nevertheless, we can try to limit DT|* beginning with a drastic move: in order to avoid the paradox, the interaction

[243] See Horwich 1998b.

[244] Some authors accept the inconsistencies, see Beall and Armour-Garb 2001, 2006 and Armour-Garb 2004.

between negation and the truth predicate must be prevented. Thus, our new truth theory should consist of the T-sentences of the form:

$T(\ulcorner\varphi\urcorner)\leftrightarrow\varphi$

where φ is a sentence in L_T in which "T" does not occur negated. Actually, even more can be conceded. Since some negations are (in classical logic) innocuous. "T" can be allowed to occur as long as it occurs only *positively*, so that it does not occur after an odd number of negations. Call the obtained theory DT+ (DT+ has full induction in L_T, we use the label "DT|+" for the theory with induction restricted to L_{PA}). Putting the lack of expandability of DT aside, that a deflationist should accept this theory is a reasonable expectation. DT+ is based only on T-sentences that appear to be completely respectable. It does not seem that there are reasons to prevent the theory from containing biconditionals like: $T(\ulcorner\neg\neg TS0=0\urcorner)\leftrightarrow\neg\neg T(\ulcorner S0=0\urcorner)$. Certainly DT+ cannot be considered a final adequate theory, but, although it is not clear how to deal with the paradox and to limit the set of T-sentences, DT+ is a good improvement. Although it is not clear what T-sentences the final theory should contain, we can say that it should contain at least those in DT+. DT+ should be regarded as a subtheory of every good deflationary theory. Again, however, problems are in the neighbourhood. Volker Halbach[245] has investigated a version of DT+, showing that in it the truth predicate of the axiomatic theory KF is definable. KF[246] is the axiomatic version of the truth theory inspired by the work of Saul Kripke and it is one of the strongest[247] truth theories currently on the market. KF, for

[245] Halbach 2009.

[246] There are different axiomatic versions of the Kripkian theory. "KF" is usually used to refer to (one of) the versions elaborated by Solomon Feferman.

[247] KF has the same strength of Ramified Analysis up to the ordinal ε_0, which is equivalent to Tarskian truth iterated ε_0 times.

instance, is much stronger than T(PA). The version of DT+ investigated by Halbach is what he calls PUTB, obtained by imposing uniformity to DT+:

PUTB: $\forall x(T(\lceil \varphi(\mathbf{x}) \rceil) \leftrightarrow \varphi(x))$

where φ is a sentence in L_T that contains the truth predicate only positively.

To give a complete description of KF, and a complete reconstruction of the work of Halbach would require too much space and lead us astray. So we limit ourselves to an exposition of the main points. We refer to Halbach (2009) for an exhaustive treatment. The first relevant result is that PUTB is consistent (Halbach shows this by constructing a model for the theory). Hence the idea of limiting the truth predicate only in positive occurrences to avoid the paradox is indeed an effective choice. The most important results, however, involve the relation between PUTB and KF. First of all, Halbach shows that PUTB can define the truth predicate of KF. This (and the fact that PUTB is a subtheory of KF) has the immediate consequence that both theories have the same arithmetical content: KF and PUTB prove the same sentences in L_{PA}. This has a great relevance for conservativeness. KF, in fact, is not conservative over PA: it proves sentences in L_{PA} that cannot be proved in PA alone (for instance Con_{PA}). Since KF and PUTB prove the same sentences in L_{PA}[248], it follows that PUTB is not conservative over PA either. This does not mean that KF and PUTB are equivalent theories, though. In fact Halbach proves that PUTB is a proper subtheory of KF. Such a result is interesting because the weakness of PUTB is close to that of DT. PUTB cannot prove generalization like Gen or more general compositional principles. This means that PUTB can define a truth predicate that is a compositional predicate, but the truth predicate of PUTB is not a compositional predicate. These results are important

[248] Halbach 2009, p. 4.

because they show that a deflationist can strengthen her theory without giving up pure disquotational axioms. The problem is that, at the same time, such disquotational axioms are much less innocent than they seemed to be. This is surprising: it suffices to allow a positive iteration of the truth predicate to lose conservativeness and to get a very strong theory.

A first conclusion we can draw is that the conservativeness of some simplified theory based on T-sentences is not due to the disquotational nature of T-sentences, but to the simplification imposed to make our job easier. Conservativeness does not depend on the nature of T-sentences; it depends on the big simplification we made. The premise from which the argument from conservativeness started -the essential agreement between deflationism and conservativeness- was just an illusion. If we had considered PUTB the strategy would have been unviable from the start. Clearly, a deflationist could refuse to adopt a theory like PUTB, sticking just with local disquotational axioms with restricted induction, opting for a theory like DT|+. Although Halbach conjectures that DT+ (and then DT|+) to be conservative over PA, this fact is unknown so far. Even conceding the legitimacy of this move, however, a deflationist cannot go very far along this way either. The overall general complexity and strength of T-sentences must be eventually faced.

INSIDE THE LABYRINTH OF T-SENTENCES

That DT|* is not conservative is something we might not be worried about, since it is well known that the liar paradox is not easy to solve. Exactly for this reason we have only considered simplified versions. However, we may also check what happens if our theory includes as many T-sentences

as possible, just avoiding inconsistencies, namely a theory $DT|^{MAX}$ consisting in a maximal consistent set of T-sentences. It seems undeniable, at least *prima facie*, that a deflationist should accept a theory like $DT|^{MAX}$, and that, perhaps, this is exactly the theory she has in mind. That deflationism might take the form of $DT|^{MAX}$ is confirmed and explicitly accepted by Paul Horwich[249]. His attitude towards the liar paradox is that of giving up DT|*, letting logicians to search for a solution to the Liar, and to stick with the set of T-sentences that are safe in the meantime. Unfortunately, also $DT|^{MAX}$, as McGee has shown in a brief and surprising article[250], has unexpected consequences.McGee investigates what happens if we look for the biggest number of T-sentences with the only purpose of avoiding inconsistencies without any further general or philosophical consideration leading us. The idea is that of reducing DT|* in the minimal way able to avoid the inconsistency, so that our set of T-sentences will be both maximal and consistent. McGee proves his result generally considering an arithmetical theory S which implies the minimal arithmetical theory R, namely Robinson's arithmetic. R is a subtheory of PA and we can (very roughly) think it to be just as PA without the induction schema[251].

6. 13 Theorem:

Let Δ be an S-consistent set of sentences of L_S. Then there is a set of T-sentences Γ such that

i. all the members of Δ are S-entailed by Γ,

ii. Γ is S-consistent,

[249] Horwich 1998b.

[250] McGee 1992.

[251] Strictly speaking, this is a mistake, since without the induction schema new axioms are needed. About R (or the analogue theory Q) see Boolos, Burgess and Jeffrey 2007.

iii. any set of T-sentences which properly includes Γ is S-inconsistent, and

iv. Γ ∪ R is a complete first-order theory.

The point iii. means that no new T-sentence can be added to Γ without making the set inconsistent. iii. is just states that Γ is a maximal consistent set of T-sentences, (it is our $DT|^{MAX}$). The proof is very short and it is worth giving it a look.

Proof:

i. Use Gödel's self-referential lemma to find, for each sentence φ, a sentence P_φ such that $P_\varphi \leftrightarrow (\varphi \leftrightarrow T(\ulcorner P_\varphi \urcorner))$ is a theorem of R. It follows by logic that $\varphi \leftrightarrow (P_\varphi \leftrightarrow T(\ulcorner P_\varphi \urcorner))$ is a theorem of R (where "T" is a new symbol added to the language of R).

ii. Since Δ is S-consistent, the set of all biconditionals $\psi \leftrightarrow (P_\psi \leftrightarrow T(\ulcorner P_\psi \urcorner))$ with ψ in Δ is S-consistent.

iii. By Zorn's lemma, we find a maximal S-consistent set Γ of T-sentences, which includes all the biconditionals $(P_\psi \leftrightarrow T(\ulcorner P_\psi \urcorner))$ with ψ in Δ.

iv. Take any sentence ϕ. Because Γ is S-consistent, either Γ ∪ PA ∪ {ϕ} is consistent or Γ ∪ PA ∪ {¬ϕ} is consistent. So either Γ ∪ PA ∪{$(P_\phi \leftrightarrow T(\ulcorner P_\phi \urcorner))$} is S-consistent or Γ ∪ PA ∪ {$(P_{\neg\phi} \leftrightarrow T(\ulcorner P_{\neg\phi} \urcorner)$}is consistent. It follows by iii. that either $(P_\phi \leftrightarrow T(\ulcorner P_\phi \urcorner)$ is in Γ or $(P_{\neg\phi} \leftrightarrow T(\ulcorner P_{\neg\phi} \urcorner)$ is in Γ.

McGee[252] considers the extreme case where Δ is a complete theory and "T" is treated without any intention to consider it as representation of the set of truths (its extension

[252] McGee 1992, p. 238.

could be the simple empty set, that of odd numbers or even that of arithmetical falsehoods). Even so, it turns out that Δ entails a lot of T-sentences. So many, in fact, that if we were to construct a different theory with the conscious purpose that "T" should stand for the ordinary notion of truth, inevitably there would be some T-sentences that the new theory would falsify but Γ would make true. Any complete, consistent extension (no matter how unlikely it is as a theory of truth) will entail a maximal S-consistent set of T-sentences. This means that there are many maximal consistent sets of T-sentences incompatible with each other. Among these there are some, for instance, that S-entail 2+2=4 and others that S-entail 2+2=5. In general for every sentence φ there is at least a maximal consistent set of T-sentences that S-entails φ and another that S-entails ¬φ.

Another interesting fact is the following: "If φ is grounded, in Kripke's sense, there will be some stage α in Kripke's construction at which φ has been declared either true or false but T"(⌈φ⌉) has not yet been declared either true or false. By setting the extension of "T" equal either to $S_{1,\alpha}$ or to the complement $S_{2,\alpha}$ we get a classical model of S in which φ ↔ T(⌈φ⌉) is false, and so we get a maximal S consistent set of T-sentences which excludes φ ↔ T(⌈φ⌉). Thus we find, surprisingly, that there are no grounded sentences among the sentences φ such that φ ↔ T(⌈φ⌉) is in every maximal S-consistent set of instances of (T). Instead, the only sentences φ such that φ ↔ T(⌈φ⌉) is in every maximal S-consistent set of T-sentences will be ungrounded sentences like "This sentence is true" that assert their own truth"[253].

Using McGee's tools, Christopher Gauker[254] has shown that a deflationist cannot restrict the sets of T-sentences

[253] McGee 1992, pp. 238-239.
[254] Gauker 2001.

by trying to get just those T-sentences that do not entail arithmetical falsehoods, like 2+2=5[255]. Gauker also gives a more explicit reconstruction of the technique that McGee used to prove that every sentence has a materially equivalent T-sentence. Take an arbitrary sentence of our natural language, "snow is white", for instance. We can add to our language a new individual constant τ that denotes the sentence "snow is white if and only if τ is true"

(τ) "snow is white if and only if τ is true."

Then reason as follows:

1. (snow is white if and only if τ is true) if and only if ("snow is white" is true if and only if τ is true)

(by the meaning of "if and only if")

2. (snow is white if and only if τ is true) if and only if ("snow is white" is true if and only if "τ is true" is true)

(from 1. by the relevant T-sentences and by substitution in the right side of " "snow is white" " with " "snow is white" is true" and "τ is true " with " "τ is true"is true").

3. ("snow is white" is true if and only if "τ is true" is true) if and only if (snow is white if and only if τ is true)

(from 2. by commutativity of if and only if)

4. ("snow is white" is true) if and only if (("τ is true" is true) if and only if (snow is white if and only if τ is true))

(from 3. by associativity of if and only if)

5. ("snow is white" is true) if and only if (("τ is true" is true) if and only if ("snow is white if and only if τ is true"is true)

(from 4. by the relevant T-sentences and by substitution

[255] But see Raatikainen 2002 about Gauker's position.

in the right side of "snow is white if and only if τ is true" with ""snow is white if and only if τ is true" is true".)

6. "snow is white" is true if and only if ("τ is true" is true if and only if τ is true)

(from 1. and 5. by the law of identity and by substitution in the right side of 5. of ""snow is white if and only if τ is true"" with "τ").

Point 6. gives us a T-sentence that is materially equivalent to the sentence "snow is white". Assuming 1. Gauker shows that this argument can be given even without introducing new constants or assuming dubious identities. Gödel's diagonal lemma (used by McGee too) states that, under certain conditions, for every formula F(x), with x as a unique free variable, there exists a sentence φ such that φ is true if and only if F(⌈φ⌉) is true. Assume that the language contains its own diagonal predicate D, so that the expression D(a,b) means that the sentence that ⌈a⌉ denotes is the diagonal of the sentence that ⌈b⌉ denotes. By the diagonal lemma, from the formula F(x) the sentence ∃y(D(y,⌈∃y(Dyx ∧ F(y))⌉) ∧ F(y)) can be obtained. This sentence is materially equivalent to F(⌈∃y(D(y,⌈∃y(Dyx ∧ F(y))⌉) ∧ F(y))⌉ (in virtue of the meaning of D, the first of these two sentences is true if and only if ∃y(y= ⌈∃y(D(y,⌈∃y(D(y,x) ∧ F(y))⌉) ∧F(y))⌉∧ F(y)) is true, which is so if an only if the second sentence is true)[256]. It can then be shown that every sentence of the language is materially equivalent to some T-sentence. Let S be any sentence of our language, and consider the formula:

S if and only if y is true

If this formula is substituted for F(y) in the previous formula ∃y(D(y,⌈∃y(Dyx ∧ F(y))⌉) ∧ F(y)), the following is obtained:

[256] Gauker 2001.

∃y(D(y, ⌈(∃y(Dyx ∧ S if and only if y is true)⌉) ∧ S if and only if y is true)

Let us shorten this last sentence with P. Since P is the Gödel-sentence obtained from "S if and only if y is true", "P" is true if and only if the sentence "S if and only if "P" is true" is true. We can now reason:

1.("P" is true) if and only if ("(S if and only if "P" is true)" is true); (by the argument just given)

2. ("(S if and only if "P" is true)"is true) if and only if ("S"is true if and only if "("P" is true)" is true); (by the meaning of "if and only if")

3. ("P" is true) if and only if ("S" is true if and only if "("P" is true)"is true); (from 1. and 2. by the transitivity of if and only if)

4. ("S" is true) if and only if ("("P" is true)" is true if and only if "P" is true); (from 3. by the commutativity and associativity of if and only if)

5. ("("P" is true)"is true if and only if "P" is true) if and only if ("("P" is true if and only if P)" is true); (by the meaning of if and only if)

6. ("S" is true) if and only if ("("P" is true if and only if P)" is true) (from 4. and 5, by transitivity).

7. S if and only if ("P"is true if and only if P).

Where the right side of 7. is just a T-sentence materially equivalent to S. If S is our sentence "snow is white", we get that its materially equivalent T-sentence is:

[("∃y(D(y,"∃y(D(y,x) ∧ "snow is white" is true if and only if y is true")) ∧ "snow is white " is true if and only if y is true)"is true] if and only if [∃y(D(y,"∃y(Dyx ∧ "snow is white" is true if and only if y is true)") ∧ "snow is white" is true if and only if y is true)]. Where P is the formula: ∃y(D(y,"∃y(D(y,x) ∧ "snow is white" is true if and only if y is true)") ∧ "snow is white " is true if and only if y is true).

A very complicated result indeed.

Many and important consequences for deflationism

follow from McGee's results and his construction. For instance, it follows (from point iv. of the theorem) that, if a deflationary theory consists in a maximal and S-consistent set of T-sentences, then such a theory is not axiomatizable[257]. Maximality and consistency are not enough to give us a unique set of T-sentences. What we get, instead, is a great number of candidates. Among these candidates, there are some that, although not paradoxical, can prove any sort of sentences (even false ones). To fix this, a natural temptation could be to restrict our theory to the maximal and S-consistent set that entails just truths (meaning the sentence true in the standard model $\mathbb{N}$). Gauker, however, has shown that this way is impracticable: if we could individuate such a set, we would be able also to enumerate all arithmetical truths, against Gödel's theorem.

To spot a suitable set of T-sentences, one could try to impose some syntactic restrictions. Schindler[258] has proved that in this way handleable versions of consistent theories of (uniform) T-sentences can be elaborated. However, on the one hand, from a disquotational point of view, the philosophical motivation behind such sets of T-sentences is unclear. On the other hand, very strong theories are obtained in this way[259], so they hardly represent an improvement on the conservativity side of the story. A way to correct the former defect might be that of replacing a syntactic restriction with a semantically inspired one. An appealing strategy could be that of admitting only biconditionals for grounded sentences in the sense of Kripke. However, this option is hardly viable, since the set of grounded sentences is very complex[260]

[257] The point iv. of the theorem 6.13 states that $\Gamma \cup R$ is a complete first-order theory. (Where Γ is our maximal consistent set of T-sentences). It follows from Gödel's theorem that if such a theory, including the arithmetic theory R, is complete, it cannot be axiomatizable.

[258] Schindler 2015.

[259] They can be arithmetically as strong as the theory Z_2^-.

[260] It is a Π^1_1 set.

and fails to provide a basis for an axiomatizable theory. Moreover, a great proof theoretic strength can be expected as a consequence of such a mathematical complexity.

FOLLOWING CONSERVATIVENESS OUT OF THE LABYRINTH

McGee's construction shows that, for every sentence φ, it is possible to get a T-sentence τ that is materially equivalent to φ. This means that, if a deflationary theory contains such a T-sentence τ, the theory also entails φ. Consider then the base theory PA, and let φ be a sentence in L_{PA} not provable in PA. Since it is possible to construct a T-sentence τ which entails φ, if our deflationary theory $DT|^{MAX}$ contains τ, $DT|^{MAX}$ entails φ too. Given that φ is a sentence that is not provable in PA, $DT|^{MAX}$ is not conservative over PA. If we want to respect a conservativeness requirement, the deflationary theory cannot contain any T-sentence equivalent to a L_{PA}-sentence not already provable in PA. What we should aim at, then, is not only a PA-consistent and maximal set of T-sentences. We should search for a set of T-sentences that is conservative over PA. Cieslinski[261] has studied such an option, showing that: 1. the addition of a conservativeness requirement does not suffice to get a good restriction of the number of the candidates to be a deflationary theory, and 2. the conservative sets are not axiomatizable.

6.14 Theorem:

Let Γ be a conservative extension of PA in L_T; then there is a theory Γ' in L_T that includes Γ and it is a maximal conservative extension of PA.

[261] Cieslinski 2007.

Obviously, since the predicate "T" must stands for the truth predicate, Γ must prove all sentences of the form $T(\ulcorner\varphi\urcorner)\leftrightarrow\varphi$ for every φ in L_{PA}, so that Γ includes DT|. As we know, DT| is conservative over PA.

6.15 Theorem:

Let Γ be PA ∪ DT|. Then Γ has a continuum many maximal conservative extensions.

6.16 Theorem:

Let Γ be a conservative extension of PA in L_T such that:
i. Γ includes DT|
ii. Γ is axiomatizable
then there is a sentence ψ in L_T, such that Γ does not prove ψ and T ∪ $\{\psi\}$ is a conservative extension of PA.

This means that if our theory Γ is conservative over PA, it includes DT|, and it is axiomatizable, then it is not maximal.

If these results cannot solve the problems raised above, nevertheless they seem reassuring with regard to conservativeness. A deflationist has the chance to advocate a theory like $DT|^{MAX\text{-}CONS}$, which represents a maximal and conservative set of T-sentences. The problem of picking out a single set from the great number of candidates is not something we should be worried about. Indeed, that deflationists can choose among many sets might be considered a perk: not only is it possible to combine T-sentences in such a way that we can preserve conservativeness; we can do that in many different ways too.

A FINAL PROBLEM

Unfortunately, the idea of limiting the set of T-sentences to avoid all sorts of complication leads to a deep worry for the deflationist. According to deflationism, the truth predicate serves, by T-sentences, the purpose of expressing blind truth ascription like "everything the Pope says is true". Suppose now that in order to avoid contradictions we had to limit our theory excluding all T-sentences leading to paradoxes and, similarly, all T-sentences materially equivalent to sentences in the language of the base theory that are not already provable in such a base theory[262]. Consider, for instance, the sentence "PA is consistent", and construct the biconditional materially equivalent to it: [("∃y(D(y,"∃y(D(y,x) ∧ PA is consistent if and only if y is true") ∧ PA is consistent if and only if y is true)")is true] if and only if [∃yD(y, "∃y(Dyx ∧ PA is consistent if and only if y is true") ∧ PA is consistent if and only if y is true)]. Let ψ be the sentence "[∃y(D(y,"∃y(Dyx ∧ PA is consistent if and only if y is true") ∧ PA is consistent if and only if y is true)]". The T-sentence in question then is " "ψ" is true if and only if ψ". Since we know that "PA is consistent" is not provable in PA, our theory cannot contain " "ψ" is true if and only if ψ". Otherwise our theory would be able also to prove "PA is consistent" and it would not be conservative over PA. Thus, if the Pope was to assert ψ, our truth ascription "something the Pope says is true" could not cover it, because the T-sentence ""ψ" is true if and only if ψ" is not available. At this point a deflationist can only hope that the Pope does not know enough logic to ever assert a complicated sentence like ψ.

The point is general. By embracing a set in $DT|^{MAX\text{-}CONS}$, it is likely that a deflationist could be unable to make sense of claims like "everything Kripke says in "Outline of a Theory

[262] According to theorem 6.13.

of Truth" is true", or "the sentence Eubulides is famous for is not true" or "something McGee said in his article on maximal consistent sets of T-sentences is true". Not only is there a problem in the idea of restricting the set of T-sentences to avoid paradoxes, if we have to exclude also T-sentences that are materially equivalent to those sentences in the language of the base theory that are not provable in it, the problem becomes measureless big: we have to reject an infinite number of different T-sentences.

FINAL REMARKS

The conservativeness requirement has revealed itself to be compatible with the idea of a truth theory based on T-sentences. Such a result, however, has risen hard worries. First of all, a deflationist needs to argue against the acceptability of apparently good theories (as T(PA)|, UDT, PUTB and DT). Second, even if she focuses on theories based on simple T-sentences and restricted induction (inspired by DT|), she has to give up so many T-sentences that the result is a Pyrrhic victory. In particular, $DT|^{\text{MAX-CONS}}$ (no matter how it is individuated) does not contain any T-sentence materially equivalent to sentences in L_{PA} that are not provable in PA. Whether giving up an infinite number of T-sentences could be acceptable to deflationism, however, is questionable. The fact that T-sentences have so many problems, beside the liar, and that many of them conflict with one fundamental thesis of deflationism itself (the insubstantiality of truth) is enough to question the supposed innocence of their status. The T-sentences yielded in McGee's construction have a rather complicated structure and their formulation is quite laborious. A deflationist could perhaps blame this: after all, the transparency deflationists have in mind is not exhibited in those cases. It is clear, though, that this reply would be

unsatisfactory. On the one hand, the proposal could be read as an admission of guilt: many T-sentences are not as innocent as deflationists claimed. On the other hand, if McGee's T-sentences were rejected because they are too elaborated, we should eliminate a great number of them and this would make the problem even worse.

In this chapter we did not mean to show that a theory based only on T-sentences and conservative at the same time is not available to deflationism. Indeed maximal and consistent sets of T-sentences do exist. What we aimed at, instead, is collecting several non trivial or not well known results to show that the agreement between T-sentences and conservativeness is more complicated and hard to realize than usually taken to be. Complications are so many that a deflationist is forced into a tortuous path. The idea of the centrality of T-sentences and of a conservativeness requirement can hardly be combined. In order to keep them together a deflationist is compelled into uncomfortable positions over many problems, and she has to give up an infinite number of T-sentences, sacrificing the most authentic inspiration of deflationism. The contrast between the conservativeness requirement on the one hand, and T-sentences on the other hand emerged. If deflationism wants to stick with the former it must be ready to sacrifice the latter with no mercy.

Moreover, if a deflationist renounces a big number of T-sentences, problems with the logical function of the truth predicate immediately arrive. Given that the logical function of the truth predicate is the other characteristic mark of deflationism, it is time to ask: is the logical function of the truth predicate compatible with conservativeness? This is the topic of the next chapter..

CHAPTER SEVEN
LOGICAL FUNCTION Vs CONSERVATIVENESS

REDUNDANCY, CONSERVATIVENESS, AND DEFLATIONISM

There is a point that modern deflationists repeatedly stress, at least after Quine: a truth predicate governed by T-sentences gives us an extremely useful tool to realize certain logico-grammatical purposes. Sometimes, in fact we want to assert a sentence that, for different reasons, we cannot cite explicitly. What we can do, then, is to assert it indirectly using the truth predicate. The same happens when we cannot assert all the sentences in a certain class because they are too many. If we were to eliminate the truth predicate we would lose the ability to express commitments that we could not express another way. It is not possible to eliminate the truth predicate without an impoverishment of the *expressive* power of the language. The truth predicate is not semantically redundant.

The redundancy of the truth predicate is a thesis advocated by some early formulations of the deflationist position. According to a redundantism approach everything that can be said using the truth predicate could be said without it. On this ground it is easy to argue that the

truth predicate has no content and that there is not such a property as "being true". It is also clear that if everything that is said with the truth predicate can be said without it, the introduction of such a predicate cannot increase the *deductive* strength either. Let φ be a sentence whatsoever, and φ' its equivalent truth-free translation according to redundantism. Since everything that can be said with the truth predicate can be said also without it, then there must exist a translation Δ' in the language of the base theory that is still a proof of φ. By applying redundantism, any proof Δ of the sentence φ can be transformed into a proof Δ' of φ' in which the truth predicate never occurs. After all, if the truth predicate is redundant, it is redundant also in a proof. Redundantism is thus conservative, just because everything that can be said, or proved, can be said or proved in the base language.

Modern deflationism rejects the thesis of the expressive redundancy of the truth predicate, therefore the truth predicate cannot be considered deductively redundant in the above sense either. If there are expressions that only the truth predicate permits, automatically there is at least a sentence ψ in the language of the deflationary theory L_D that is provable in B ∪ D (where B is the base theory and D a non redundandist deflationary theory) but not translatable in L_B. The point is not completely trivial, since we could expect an intermediate case: that D gives us a theory not expressively but deductively redundant. It is quite simple to give an example, though. Let ψ be a sentence in L_D not expressible without the truth predicate: then $\psi \rightarrow \psi$ is a sentence in L_D provable (by simple logic) in B ∪ D that is not provable in the base theory B, and for which a suitable translation is not available. If modern deflationism cannot hold neither expressive nor deductive redundancy, however, it can try at least to be conservative. Indeed, conservativeness is just deductive redundancy restricted to sentences in the language of base theory.

THE LOGICAL FUNCTION OF THE TRUTH PREDICATE: A MINIMAL SENSE OF *EXPRESSING*

T-sentences, according to deflationism, allow the truth predicate to serve an important role: mimicking, by finite means, certain infinitary operations. The truth predicate enables the expression of infinite conjunctions (and disjunctions) by generalizations in which the truth predicate cannot be eliminated. Deflationists, usually, follow Quine and show that the T-sentences allow the truth predicate to have this role by permitting the use of objectual quantification in sentential positions. If this part of the proposal is rather clear, deflationists devoted much less time and patience to clarify in what sense their theory is able to *express* infinite conjunctions through these generalizations, despite the fact that this is a crucial point, and that it has been also criticized[263]. In what sense can such generalizations be considered expressing infinite conjunctions (for sake of simplicity, henceforth we will focus on conjunctions putting the case of disjunctions aside)?[264] Volker Halbach in "Disquotationalism and Infinite Conjunctions" put forward a detailed answer to this question[265].

Before discussing Halbach's proposal, some premises are in order. First of all, not every set of sentences can be (univocally) expressed with a generalization. The reason is that, given the set Σ of well formed sentences of a language, the cardinality of the set of its subsets, the powerset $P(\Sigma)$, is bigger than the cardinality of Σ. Thus we have more sets of sentences than sentences or generalizations. It follows that some set of sentences cannot be expressed by a generalization. A reasonable limitation is then narrowing our attention to those sets that are definable in the language.

[263] See Gupta 1993.
[264] See also Picollo and Schindler 2018.
[265] Halbach 1999b.

To avoid complications from paradoxes, we consider those sets of sentence in L_{PA} defined by a formula in L_{PA} and the corresponding generalizations formed in L_T[266]. Beside such practical restrictions, it is important to notice that we cannot impose other limits on what sets of sentences a deflationary truth theory should be able to express. If the truth predicate must serve its logico-grammatical purpose, then, given a set Δ of sentences in L_{PA} with the form $\delta(\ulcorner\varphi\urcorner)\rightarrow\varphi$ where $\delta(x)$ is a formula in L_{PA} defining a set of sentences in L_{PA}, the theory must provide a sentence in L_T expressing the infinite conjunction of the sentences in Δ.

Volker Halbach[267] has elaborated a notion of expressing infinite conjunctions[268] that is available also to conservative theories of truth. From our point of view the proposal is highly valuable. First of all, it is technically precise and we can evaluate it clearly. Second, it gives us what can be considered a minimal sense of "expressing an infinite conjunction". Finally, at least prima facie, it can be combined with conservative theories of truth, so that it seems to be a practicable way to join the logical function together with conservativeness. Let us analyse these three points.

---

[266] We follow Halbach 1999b.

[267] Halbach 1999b.

[268] We can exclude some other options immediately. Certainly "expressing" cannot be interpreted in the sense that the infinite conjunction must be "provable". First, this would be untenable because we do not expect a theory of truth to prove, for instance, that "everything the pope says is true" (see Halbach 2001b, p. 184), or, even worse, to prove any infinite conjunction. If we were to express in this sense the infinite conjunction saying that everything PA proves is true, we would lose conservativeness immediately. What we should demand is something else. We need, at most, that whenever a set Δ of sentences in L_{PA} with the form $\delta(\text{“}\varphi\text{”})\rightarrow\varphi$ (where $\delta(x)$ is a formula in L_{PA} defining a set of sentences in L_{PA}) is assumed, a theory of truth should be able to prove the generalization n L_T $\forall x(\delta(x)\rightarrow T(x))$. This is acceptable and very close to the adequacy requirement. However if a truth theory has such a proof strength it is not a conservative theory over PA. We can see this considering Heck's proof below.

Halbach considers a formula $\forall x(\delta(x) \rightarrow T(x))$ in L_T expressing an infinite conjunction of sentences in L_{PA} if the addition of the sentences in Δ and the addition of the formula $\forall x(\delta(x) \rightarrow T(x))$ (together with axioms for truth) have the same consequences over PA, namely the base theory. The general definition of Halbach is: "let $\delta(x)$ be a formula of L_B with exactly x free, τ the sentence $\forall x(\delta(x) \rightarrow T(x))$ and define B' as the base theory expanded by all sentences $\delta(\varphi) \rightarrow \varphi$ where φ is a sentence of L_B, then B' e $B_1 \cup \{\tau\}$ prove the same L_B-formulas"[269] (where B_1 is the base theory expanded by T-sentences). Halbach proves that if the theory of truth is DT|, generalizations like $\forall x(\delta(x) \rightarrow T(x))$ and the corresponding infinite conjunctions really have the same consequences over the base theory, and then he extends the treatment to disjunctions. We can also put the proposal explicitly in terms of conservativeness. The idea is that the addition of the infinite conjuncts and the addition of the corresponding generalization must yield equivalent extensions of the base theory. With respect to the base theory it must not make any difference if the set of the infinite conjuncts or the truth generalization is added.

7.3 Definition:

let τ be the sentence in L_T $\forall x(\delta(x) \rightarrow T(x))$, Δ a set of sentences in L_{PA} with the form $\delta(\ulcorner\varphi\urcorner) \rightarrow \varphi$ where $\delta(x)$ is a formula in L_{PA} defining a set of sentences in L_{PA}, and Γ a theory of truth (such that it can prove all the biconditionals of the form $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$ for every sentence φ in L_{PA}, that is DT| must be included in Γ). τ is said to be able to *minimally express* (or to *m-express*) the infinite conjunction of the sentences in Δ in the sense that for every sentence φ in L_{PA}

PA $\cup$ DT| $\cup\{\tau\} \vdash \varphi$

[269] Halbach 1999b, p. 13, proposition 2.

if and only if
PA ⋃ Δ ⊢ φ

Analogously, the truth theory Γ is said to be able to *m-express* the infinite conjunction of the sentences in Δ by the generalization τ.

Note that the choice of DT| is not *ad hoc*. DT| is chosen because it provides the minimal information needed to pass from an infinite conjunction in which objectual quantification over sentences is not possible to an infinite conjunction where it is possible. Accordingly, a truth theory can m-express the infinite conjunction of the elements in Δ if it includes DT| and, for every sentence φ in L_{PA}, if PA ⋃ DT| ⋃ {τ} (and not PA ⋃ Γ ⋃ {τ}!) proves φ, then PA ⋃ Δ proves φ too. The truth theory Γ must include the T-sentences in DT| but it has no other roles in such a definition. Note that, if we considered PA ⋃ T(PA) ⋃ {τ} instead of PA ⋃ DT| U{τ}), then a truth theory like T(PA), which is not conservative over PA, would not be able to m-express infinite conjunctions. Clearly, it would be unreasonable to propose a sense of expressing unavailable to stronger theories of truth like T(PA). Moreover, such a definition is available even if our theory is a conservative theory, like in the case of DT or T(PA)|. In this way we have a precise sense of "expressing" which holds also for deductively weak theories unable to prove infinite generalizations. Indeed, since a minimal constraint (by the material adequacy condition of Tarski) is that a truth theory must be able to prove all the T-sentences for L_{PA}, a truth theory must include DT|. For that reason any theory of truth has enough resources for the logical function as minimally defined in 7.3. Not only a conservative theory but every theory of truth can m-express an infinite conjunction. Even T(PA), for instance, can m-express infinite conjunctions. That T(PA) is also able to *prove* some infinite

generalizations is just something more.

Halbach's definition provides a precise and modest sense to interpret the equivalence between a generalization involving the truth predicate and the corresponding infinite conjunction. Accordingly, if a theory cannot *m-express* infinite conjunctions, it can not *express* them at all. Such a theory would be a theory unable to serve the desired logical function. The idea we started from is that in some circumstances there is an infinite number of sentences about the world (non semantical) that we are not able to handle because they are too many. The truth predicate enables us to solve this problem. This is possible, though, only if the generalization and the assumption of the corresponding set of sentences have the same effects on the base theory. The idea is: "instead of giving a list of all the infinite conjuncts, I join them together in a single generalization, just to save ink and time". If the consequences were different, we would be hardly allowed to use one instead of the other. This does not mean that the m-expression is considered as an adequate account. Rather, what we want to emphasize is that whatever sense of expression is proposed, it must include the minimal sense spelled out here. That a truth generalization expresses an infinite conjunction must at least mean that the former m-expresses the latter.

A last remark concerns the relation between m-expression and conservativeness. Here we are considering truth theories that are conservative over PA, so that to show that the conservativeness requirement is respected, we must just check whether m-expression does not bring heavy consequences. A theory like DT, for instance, is conservative and it cannot prove any infinite generalization τ. However, it can m-express infinite conjunctions through τ. To evaluate the possible agreement between m-expression and conservativeness we have to compare the part of m-expressing involving truth ($PA \cup DT|_{\cup} \{\tau\}$) with the

part truth-free (PA ⋃ Δ). The addition of a generalization τ (with the relevant T-sentences) and the addition of the set of conjuncts Δ must have the same consequences on the base theory. If this was not the case the m-expression of an infinite conjunction would make us lose conservativeness.

From the definition given above, it is easy to get a reassuring result: the logical function of the truth predicate characterized by the notion of m-expressing respects the conservativeness requirement, at least in some simple cases.

7.4 Proposition:

Let τ be the m-expression of the infinite conjunction of the elements in a set Δ of sentences in L_{PA} with the form $\delta((\ulcorner\rho\urcorner)\rightarrow\rho$ where δ(x) is a formula in L_{PA} defining a set of sentences in L_{PA},

For any sentence φ in $L_{PA.}$

If PA ⋃ DT| ⋃ {τ}⊢ φ

then

PA ⋃ Δ ⊢ φ.

(The proof follows immediately from the definition of m-expressing).

This proposition shows that the logical function of the truth predicate has no substantial consequences and a deflationist can adopt it. Note that it would be a mistake to demand conservativeness without assuming also the set Δ, namely that:

for every φ in L_{PA} if PA ⋃ DT| ⋃ {τ}⊢ φ then PA ⊢ φ.

This would mean that the m-expressed infinite conjunction should be conservative over PA. This is unjustified. If the conjunction was not conservative (for instance if it contained $Prov_{PA}(\ulcorner Con_{PA}\urcorner)\rightarrow Con_{PA}$), also the assumption of τ would not be conservative but

the responsibility would be not be on truth, but in the conjunction τ expresses.

THE LOSS OF INNOCENCE OF THE LOGICAL FUNCTION

The results above are all positive for deflationism. Thanks to the notion of m-expressing we can make sense of the idea that a truth predicate serves, by T-sentences, the logical function of expressing infinite conjunctions. This sense of expressing is arguably a minimal sense of interpreting such a logical function. Moreover, the logical function seems also compatible with conservativeness. Truth seems to be able to serve its logical role without compromising its insubstantial nature.

Unfortunately, also this time things are more complicated. Richard Heck[270] has shown that proposition 7.4 does not hold if we m-express two infinite conjunctions. Let us see Heck's case.

Consider two sentences in L_T:

τ': $\forall x(\exists n(x= \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0=S0\urcorner)\urcorner)\rightarrow T(x))$

τ'': $\forall x(T(x) \rightarrow (\forall n (x= \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0=S0\urcorner)\urcorner)\rightarrow\neg Proof_{PA}(n,\ulcorner 0=S0\urcorner))$

and let Δ' be the set of instances of τ' in L_{PA} (that is the sentences of the form: $(\exists n(\ulcorner\varphi\urcorner = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0=S0 \urcorner)\urcorner)\rightarrow\varphi$, where φ is a sentence in L_{PA}) and Δ'' the set of instances of τ'' in L_{PA} (that is the sentences of the form:

$(\varphi\rightarrow(\forall n (\ulcorner\varphi\urcorner = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0=S0\urcorner)\urcorner) \rightarrow\neg Proof_{PA}(n, \ulcorner 0=S0\urcorner)$, where φ is a sentence in L_{PA}).

It can be shown that a sentence ψ in L_{PA} exists such that

$PA \cup \Delta' \cup \Delta'' \cup DT| \cup\{\tau'\} \cup\{\tau''\}\vdash \psi$

but

$PA \cup \Delta' \cup \Delta'' \nvdash \psi$.

[270] Heck 2004, appendix.

Note[271] that the instances of τ' and τ'' (the members of Δ' and Δ'') are already provable in PA. The antecedent of τ' is, in each case, decidable, so it is refutable if false and it is provable in PA if of the appropriate form, but then so it is the consequent, since PA does prove, for *each* natural number, that it is not a proof of 0=S0. The same holds for the instances of τ'': they are decidable if φ is of the form $\neg Proof_{PA}(n,\ulcorner 0=S0\urcorner)$; if it is not, then $\forall n\ (\ulcorner\varphi\urcorner \neq \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0=S0\urcorner)\urcorner)$ is provable in PA, the consequent follows logically, and so does the relevant instance of τ''. If it is of the right form we can prove that it is so and thus we can prove:

$\forall n(\ulcorner\varphi\urcorner = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0\urcorner)\urcorner \leftrightarrow (n=\mathbf{k}))$, for some **k**. From which it follows that the consequent is equivalent to $\neg Proof_{PA}(\mathbf{k}, \ulcorner 0 = S0\urcorner)$, that, again, is provable in PA. Since the instances of τ' and τ'' are already provable in PA, it follows that PA ∪ Δ' ∪ Δ'' is a conservative extension of PA (trivially because Δ' and Δ'' are already included in PA, so they do not yield extensions at all).

PA ∪ Δ' ∪ Δ'' ∪ DT| ∪{τ'} ∪{τ''} (that is equivalent to PA ∪ DT| ∪{τ'}∪{τ''}), however, is not conservative over PA (and hence it is not conservative over PA ∪ Δ' ∪ Δ'' either). For pure logic, in fact, τ' and τ'' imply

1. $\forall x((\exists n)(x = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0\urcorner)\urcorner) \rightarrow \forall n(x = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0\urcorner)\urcorner \rightarrow \neg Proof_{PA}(n, \text{`}0 = 1\text{'}))$

which in turn implies

2. $\forall x((\forall n)(x = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0\urcorner)\urcorner \rightarrow \neg Proof_{PA}(n, \ulcorner 0 = S0\urcorner))$

and so

3. $\forall n((\exists x)(x = \ulcorner\neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0\urcorner)\urcorner \rightarrow \neg Proof_{PA}(n, \ulcorner 0 =S0\urcorner))$

but in PA it is possible to prove that there is for every n, a sentence saying that n is not (the code of) a proof of 0=S0, that is

[271] I follow Heck 2004 here.

4. $\forall n(\exists x)(x = \ulcorner \neg Proof_{PA}(\mathbf{n}, \ulcorner 0 = S0 \urcorner) \urcorner)$

and so:

5. $\forall n(\neg Proof_{PA}(n, \ulcorner 0 = S0 \urcorner))$, that is just Con_{PA}.

Therefore PA $\cup$ DT| $\cup \{\tau'\} \cup \{\tau''\}$ proves a sentence in L_{PA} that cannot be proved in PA alone, so it is not a conservative extension of PA.

The agreement between the logical function of the truth predicate and conservativeness, shown in proposition 7.4, fails for a slightly more complex case, as the one just considered. It is enough that the expressed infinite conjunctions are more than one (in the current case just two!) and conservativeness is lost. Once again, conservativeness seems due more to an imposed simplification than to the real innocence of deflationary truth.

At this point the situation for deflationism becomes very serious. Since m-expression has been recognized as a minimal sense for the logical function of the truth predicate, the logical function of truth seems incompatible with conservativeness. Before discussing the significance of this, some remarks are in order. First of all, as noted, a deflationist cannot impose restrictions over the infinite conjunctions that are expressible. If the truth predicate serves its goal, it must enable the expression of whatever infinite conjunction. Note also that Δ' and Δ'' are not problematic. Indeed, they are already included in PA. Thus, a deflationist has no reason to reject the possibility of expressing those conjunctions. Neither can she restrict the function to the expression of a single infinite conjunction. It must well be possible to express the conjunction of two different infinite conjunctions.

CONSERVATIVENESS AND REDUNDANCY

At this point a deflationist faces a dilemma: either truth is an insubstantial property or the truth predicate serves the logical function of expressing infinite conjunctions, but not both. Indeed, rejecting the logical role is the only way to save the metaphysical insubstantiality of truth. Deaflationist could follow this route by claiming that the truth predicate serves, at most, the function of expressing finite conjunctions. Accordingly, suppose that the proposed theory Γ includes T-sentences (so it includes DT|) but the expression (least of all the proof) of *any* infinite generalization τ[272] is not permitted. We know that PA can, by itself, define a partial truth predicate. For the theorem 2.2, the set Tn of (codes of) sentences in L_{PA}, with complexity less or equal to n, true in the standard model, is arithmetically definable. For any set of arithmetical sentences with finite logical complexity, there is an L_{PA}- formula that can represent its truth. What PA cannot do is amalgamate such finite truth predicates in a single formula holding universally for every arithmetical sentence. This fact, together with what proposition 2.5 establishes (that DT can prove only finite generalizations) allows to prove that DT is conservative over PA. Anytime DT proves a sentence in L_{PA}, in fact, it does that handling only a finite number of sentences, because the proof must be finite. The same happens here. As long as a finite number of sentences is considered, PA is enough. So, if our truth theory Γ expresses only (at most) finite generalizations, PA does not need such a theory, and Γ is redundant with respect

[272] Certainly we could permit the expression of some infinite generalizations (as in T(PA)|) without losing conservativeness, but this is not enough. If we maintain that the deflationary truth predicate makes sense of the logical function, then it should serve this goal every time it is needed.

to PA[273]. In this case the theory Γ is both expressively and deductively redundant (in the sense explained above). If we eliminate the possibility of expressing infinite conjunctions in Γ, allowing just explicit truth ascriptions or finite generalizations, we are brought back to the redundancy of the truth predicate. If so, the deflationary theory that really is conservative is only, at most, the redundancy theory of truth.

The moral of the story is that deflationism has to choose: either redundantism is embraced again, or deflationism is given up. Since redundantism seems to be a no longer attractive or defensible position and the logical function has been highly exalted by modern deflationism, the only possible way out seems to be giving up insubstantiality and break the chains of conservativeness. Deflationary truth reinflates.

[273] Remember that if Γ expresses infinite generalizations (even without proving them), then Γ is not deductively redundant, because it is not expressively redundant.

CHAPTER EIGHT
CONCLUSION

A typical feature of deflationism, which can be found in any of its versions, is the idea that truth has not a substantial nature. Deflationists passed from very strong positions denying the very existence of a property of truth, to more modest proposals holding that truth is a property of a very special kind: it is a logical and insubstantial property. This however, has rendered the claim more and more obscure. At the same time, deflationists cannot just abandon the insubstantiality thesis. Without it, deflationism itself would be rejected in favour of a substantialist view, maybe a primitivist one. Explaining insubstantiality with conservativeness would have represented a solution. Deflationists would have provided an elegant explanation of the insubstantiality of truth. Thus, it is not surprising that even if the critics of deflationism proposed conservativeness, the idea has been enthusiastically adopted by most deflationists too.

However, the adoption of conservativeness seems to condemn deflationism to inadequacy. This is what the argument from conservativeness claims. Deflationists have tried to escape this conclusion in many ways, working out refined solutions and arguments. The debate has certainly contributed to address and clarify many critical points. Altogether, however, it seems that deflationists have not been able to meet the challenge, so that some authors simply

proposed to give up conservativeness, implicitly admitting the end of deflationism.

In this work we have taken a step back in order to evaluate in what measure conservativeness is compatible with the other two central claims of modern deflationism: the centrality of T-sentences and the logical function of the truth predicate. The first result is that no theory of truth can respect a universal requirement of conservativeness. No truth theory including T-sentences is conservative over logic. However, if a convincing way out was found, T-sentences and conservativeness keep clashing even over a theory of syntax like PA. If we accept the conservativeness interpretation of the insubstantiality thesis, also the requirement of the expandability of models should be accepted. Every model of the base theory should be expandable to a model of a deflationary theory of truth. Unfortunately, this is not what happens for most truth theories apparently acceptable to a deflationist. Even if we focus only on a conservativeness requirement, however, other problems emerge when simplified deflationary theories are replaced with richer ones. To afford a conservative theory of truth, an infinite number of T-sentences must be rejected. While this is already a too high price to pay, with it deflationists risk losing also the other basic idea of deflationism: the logical function of the truth predicate.

Although it is not completely clear how the truth predicate can serve such a function, the notion of *m-expressing* provides a minimal sense every deflationist should be committed to. Unfortunately, again, if the notion of truth serves the logical function even in such a minimal sense, then conservativeness is lost again. A deflationist must then choose between keeping the logical function, or keeping conservativeness by going back to redundantism.

BETWEEN THE LOSS OF SUBSTANTIALITY AND THE LOSS OF CONSERVATIVENESS

According to such results, deflationism seems an untenable conception. In particular, if the insubstantiality claim cannot be held, deflationists should probably just adopt a primitivist view, incorporating in it some of the typical features of deflationism. If this was the outcome, our inquiry would not have been a waste of time. Deflationists have often advocated a methodological deflationism: a kind of research moving from a deflationist hypothesis on truth, trying to defend it as long as possible in order to clearly see what must be added or corrected. However, it is not obvious that this must be the moral of the story. Although deflationism and conservativeness do not seem compatible, we could also read the datas the other way around. After all, that by requiring conservativeness a deflationist position is made impossible seems a too strong conclusion. If deflationism is considered a possibly wrong but at least a meaningful view, such an outcome is enough to wonder whether the conservativeness requirement, as formulated, is really appropriate. Perhaps, the requirement could be reformulated in a new more acceptable manner. In the next sections we will analyse the previous results in order to sketch a new criterion of conservativeness that seems able to do justice to the insubstantiality thesis and, at the same time, be compatible with the other deflationist claims.

BACK TO THE ORIGINS

Deflationism has its historical original motivation in reflections on the relation between an explicit ascription of truth (like: "snow is white" is true) and the assertion of the sentence truth is ascribed to (snow is white). Frege and

Ramsey, for instance, emphasised and explained in different ways what seems to be an equivalence of some kind. Indeed, it is on such an alleged equivalence that the whole history of deflationism is grounded. It is such an equivalence that motivated the idea of the insubstantiality of truth. This is why we should think that truth lacks a robust nature, because truth does not seem to make much difference. Certainly, we have to clarify in which sense it does not make a difference, given that in a lot of senses it does. However, truth equivalences and truth insubstantiality seem deeply bound. This is confirmed by the fact that when the strongest kind of equivalence has been adopted (claiming that the addition of the truth predicate adds nothing at all to the content asserted) the strongest kind of insubstantiality has been endorsed too (claiming that the property of truth does not exist).

One chief way to explain the insubstantiality of truth, as we know, is in terms of conservativeness. The strategy we want to sketch now is that of trying to follow the path from the other direction, using conservativeness to clarify the equivalence between a truth ascription to a sentence and the assertion of the sentence itself. The hope is that in this way a better criterion of insubstantiality can also be obtained. If we use the notion of conservativeness, a simple way to explain the equivalence in question can be the following.

> Let φ be a sentence in L_{PA}, the sentence $T(\ulcorner\varphi\urcorner)$ and the sentence φ can be considered *equivalent* in the sense that their addition has the same effect on the base theory.

The reflections that can be made on this regard are close to those used to motivate the notion of m-expressing. Clearly, if we add the sentence $T(\ulcorner\varphi\urcorner)$ we need to add, at least, the corresponding axiom that says that "T" represents

the (deflationary) truth predicate for φ, which is the T-sentence[274] $T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi$. A little more formally then we get:

8.1 Definition (Conservative Equivalence):

Let φ be a sentence in L_{PA}, the sentence $T(\ulcorner\varphi\urcorner)$ in L_T is said to be *conservatively equivalent* (or *c-equivalent*) to φ if, for every sentence ψ in L_{PA}

$PA \cup \{\varphi\} \vDash \psi$
if and only if
$PA \cup \{T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi\} \cup \{T(\ulcorner\varphi\urcorner)\} \vDash \psi$.

Conservative equivalence can then be straightforwardly used to obtain a notion of local conservativeness, as follows:

8.2 Definition (Local Conservativeness):

The set $\{T(\ulcorner\varphi\urcorner)\} \cup \{T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi\}$ is *locally conservative* if, for any sentence φ, ψ in L_{PA},
if $PA \cup \{T(\ulcorner\varphi\urcorner) \leftrightarrow \varphi\} \cup \{T(\ulcorner\varphi\urcorner)\} \vDash \psi$
then
$PA \cup \{\varphi\} \vDash \psi$.

Now the idea is that the insubstantiality of deflationary truth should be explained leveraging on such a notion of local conservativeness, rather than in terms of the standard global conservativeness. Namely, deflationary truth is insubstantial if an explicit truth ascription, together with the relevant T-sentence that makes (deflationary) sense to such an ascription (plus the base theory PA), is conservative over the simple sentence truth is ascribed to (plus the base theory

[274] We add T-sentences because we are considering DT. If we were dealing with different theories of truth we should add axioms telling us that "T" represents the truth predicate according to that theory.

PA). Naturally, it would not be possible to ask simply that the addition of $T(\ulcorner\varphi\urcorner)$ (with the corresponding T-sentence) should have no substantial consequence, because in this way we would not distinguish between the consequences of φ and the consequence of the ascription of truth to φ. While the precise technical elaboration and generalization of the proposal is left to another work, here we prefer to spend some words to motivate such an idea, arguing that the strategy is indeed worth pursuing.

First of all, the difference with respect to standard conservativeness should be apparent. In the standard case the addition of the whole truth theory should be conservative, while local conservativeness is a weaker request. Nevertheless it is reasonable to take this new notion as a good explanation of the insubstantiality of truth. Truth could be hardly considered insubstantial, if it did not satisfy such a requirement. There are also reasons to think that local conservativeness is a sufficient condition for insubstantiality. If the ascription of truth to a sentence does not imply, in the base non semantic language, nothing more than we would have obtained adding that very sentence, then the truth ascription has not ascribed anything substantial to that sentence.

It is important to notice that local conservativeness demands conservativeness just for explicit ascriptions, not for blind ascriptions. Take the sentence "grass is green": truth is not substantial because "grass is green" and ""grass is green" is true" have the same effects on the base theory. Now suppose that Karl yesterday had said (only) "grass is green". The sentence "what Karl said yesterday is true" might be thought to be c-equivalent to "grass is green", since what Karl said yesterday is just "grass is green" and, by substitution, we would get exactly "grass is green" is true". Then "grass is green" and "what Karl said yesterday is true" should yield equivalent extensions. However, this is not

right. According to the classical reconstruction, "what Karl said yesterday is true" cannot be interpreted as an atomic sentence, but as a quantified one like "for every sentence x, if yesterday Karl said x, then x is true". This should be interpreted as expressing the usual infinite conjunction:

(if yesterday Karl said "grass is green" then "grass is green" is true) and (if yesterday Karl said "snow is white" then "snow is white" is true) and …

Local conservativeness should then be required for each single conjunct truth is explicitly ascribed to, namely, for "grass is green", "snow is white", and so on. We could certainly demand "what Karl said yesterday is true" to respect local conservativeness also with regard to the relevant infinite conjunction. In other words, we could demand it to m-express the infinite conjunction. However, such a request (notice!) would have nothing to do with the substantiality of truth, but only with the thesis that the truth predicate serves its logical function. If the generalization did not m-express the relevant conjunction we would be allowed to conclude that the former does not express the latter, and that, probably, the thesis of the logical function is wrong. However, no substantiality of truth would follow from this. Truth could be an insubstantial property unable to give any particular logical tool (as claimed by redundandists). To hold the insubstantiality of truth, local conservativeness at the level of explicit ascriptions is enough. Indeed, the same considerations hold also if many generalizations are considered at once, like in Heck's case. Here probably lies the main mistake of the argument from conservativeness proposed by Shapiro and Ketland: the difference, for the insubstantiality thesis, between blind and explicit truth ascriptions has been neglected.

VIRTUES AND SINS OF SENTENTIAL QUANTIFICATION

Consider the way in which the two generalizations τ' and τ'' (seen in the previous chapter in Heck's proof) show the non conservativeness of the logical function intended as m-expression of infinite conjunctions. From τ' and τ'' we get, by simple logic, without using T-sentences or other principles of truth, the sentence in L_{PA} (in which "T" does not occur):

1. $\forall x((\exists n)(x=\lceil\neg Proof_{PA}(\mathbf{n}, \lceil 0 = S0\rceil)\rceil)\rightarrow\forall n(x=\lceil\neg Proof_{PA}(\mathbf{n},\lceil 0 = S0\rceil)\rceil\rightarrow \neg Proof_{PA}(n, \lceil 0 = S0\rceil))$

From which Con_{PA} is proved using only resources available in PA. As Heck remarked, such a result would be derivable also without T-sentences. Consider gradually what happens here. If Δ' and Δ'' are added to PA, (which are the infinite instances of τ' and τ''), a conservative extension of PA is obtained, since no sentence in L_{PA} not already derivable in PA can be derived. Indeed, Δ' and Δ'' are already in PA. Now suppose that L_{PA} is extended to the language L_T adding a new symbol "T" and that PA is extended by the set T-Δ' of all truth-instances of τ', namely all sentences:

$\exists n(\lceil\varphi\rceil=\lceil\neg Proof_{PA}(\mathbf{n},\lceil 0=S0\rceil)\rceil)\rightarrow T(\lceil\varphi\rceil))$

for every sentence φ in L_{PA}. Clearly, to make sense of such truth ascriptions in T-Δ', also a T-sentence for each φ to which truth is ascribed must be added. In other words, DT| must be added. In this way the extension PA ⋃ T-Δ' ⋃ DT| is obtained.(remember that Δ is already included in PA). PA ⋃ T-Δ' ⋃ DT| is still a conservative extension of PA (and the same holds if we replace T-Δ' with T-Δ''). Actually PA ⋃ T-Δ' ⋃ DT| is equivalent to PA ⋃ DT|, as it is easy to verify. The addition of the (explicit) ascriptions of truth in our derivation, then, has no substantial consequences. Now add only the generalization τ' (or only τ'') to PA ⋃ DT| and note that we still get a conservative extension of PA.

Indeed, only when eventually both generalizations τ' and τ'' are added to PA ⋃ DT| substantial consequences are derived and conservativeness is lost. So, 1. until such generalizations are introduced separately we respect conservativeness, and the same happens if 2. we add τ' and τ'' but prevent the generalizations from interacting with each other. Exactly when we allow this interaction, conservativeness is lost. What makes us lose conservativeness and what has substantial consequences, hence, are not the explicit truth ascriptions, nor the chance of having generalizations expressing (or better m-expressing) infinite conjunctions, but only the chance of operating deductions using those generalizations. Moreover, in such deductions the notion of truth does not do any job and what only matters is the syntactic form of τ' and τ''. Truth is only needed to form τ' and τ'', but then anything can be forgotten about truth. We just need to know that "T" is a predicate, regardless of what it means, and make an exercise of basic logic applying rules of connectives and quantifiers. What makes Con_{PA} derivable and lose conservativeness is having two generalized sentences instead of infinite instances. It is then reasonable to conclude that it is the tool of generalization that has substantial consequences, as Halbach already suggested. Generalization over sentences enables syntactic operations, which are the normal logical rules governing connectives and quantifiers, that could not be not applied over infinite sets of sentences. Without quantification over sentences, we must handle infinite sets of sentences (for example the set of all instances of τ' and τ'') that cannot be combined to draw consequences. If we wanted to do that (without allowing generalizations like τ' and τ'') we should be able to handle infinitary operations like infinite conjunctions. To do that, PA should be enriched with an infinitary logic[275].

[275] On infinitary logic see Bell 2016.

At this point, even if one was convinced of the thesis that quantification over sentences gives us a powerful tool and that the loss of conservativeness should be blamed on it, she might still think that the notion of truth is also responsible, since it is the truth predicate that provides such a tool. After all, it is thanks to truth, and the sentential quantification it permits, that we lose conservativeness. Although truth has no role in the derivation of Con_{PA} from τ' and τ'', it has a role in obtaining τ' and τ''. So the argument is somehow grounded on truth. Without the truth predicate the deduction could not even start.

This objection, however, neglects an important point: above we have proposed that what matters for the insubstantiality of truth is not a generic form of conservativeness, but a sort of local conservativeness only involving explicit truth ascriptions. Accordingly, it can be shown that truth is insubstantial in a precise sense, since local conservativeness is arguably respected in the derivation. What happens when blind ascriptions, like τ' and τ'', are involved, does not matter for insubstantiality. Truth is only needed to form τ' and τ'', and here truth is insubstantial for local conservativeness is arguably respected. It is just when the two generalizations (τ' and τ'') are involved that conservativeness is lost. In a nutshell: conservativeness is lost but local conservativeness is not. Therefore insubstantiality of truth can be combined with the logical function.

The moral of this analysis is that quantification is the real responsible. A confirmation can be found in those deflationary approaches that replace truth with propositional quantifiers. Similar non conservativeness results are obtained in these cases but, since truth is not treated as a predicate, we cannot conclude that it stands for a substantial property. At most we could conclude that propositional quantification is substantial. According to

modern deflationism the point of having a truth predicate is just the chance of avoiding the (possibly new) machinery for propositional quantification (such as substitutional quantifiers). Thanks to truth we can use our good old fashioned objectual quantification. However, the moral of the story is the same. Truth enables us to quantify in sentence positions and it is this type of quantification that has great consequences. Deflationary truth is an innocent and thin part of a substantial and thick quantifier.

GOD DOES NOT CARE ABOUT TRUTH: DEFLATIONISM AND INFINITARY LOGIC

Reflecting on the argument above an objection is spontaneous. Admit that our move works: if it is quantification what makes us lose conservativeness, have we just proved that quantification is substantial? If this was the result we would jump out of the frying pan into the fire: we would have saved the innocence of truth at the cost of making quantification metaphysically robust. Is not this a paradoxical result?

Reasons to think otherwise are available. According to deflationism the utility truth is that it allows to mimick with finite means infinitary operations, like infinite conjunctions and disjunctions. We cannot assert, for example, all the theorems of PA without saying something like "everything PA proves is true". We might choose to list and to assert those theorems one by one, avoiding the truth predicate, but this would demand an infinite amount of time. This is why we need the device of the truth predicate. Possibly a God, who arguably would have no problems of time, could do the same without the truth predicate. She could use infinite conjunctions and disjunctions directly. She could even combine infinite conjunctions to draw conclusions. She could

reason with an infinitary logic. Note that infinitary logic is very powerful. For instance it is possible to give an infinitary version of PA able to characterize the standard model $\mathbb{N}$. So its addition to PA would yield a non conservative extension. Is this a proof that the logical connectives of an infinitary logic are metaphysically robust? Maybe. The question is an interesting topic for the philosophy of logic. What we want to suggest, here, is just that, if those infinitary operations are thought to be substantial in some sense, then it should not be surprising that the quantification mimicking it, obtained by the truth predicate, is substantial as well.

This manoeuvre replies to the argument of Shapiro and Ketland in a new way. The conservativeness requirement has been reformulated in a local way, showing that it can be used in a way making better sense of insubstantiality, and giving room to satisfy the adequacy requirement at the same time.

Deflationism goes on.

BIBLIOGRAPHY

Armour-Garb, B. 2001: "Deflationism and the Meaningless Strategy", *Analysis*, 61.4, pp. 280-89.

Armour-Garb, B. 2004: "Minimalism, the Generalization Problem, and the Liar", *Synthese*, 39, pp. 491-512. Asay, J. 2014; 'Against Truth'. *Erkenntnis* 79 (1):147-164.

Azzouni, J. 1999: "Comments on Shapiro" *The Journal of Philosophy*, Vol. 96, No. 10 , pp. 541-544.

Bar-On, D., Horisk, C. and Lycan, W.G. 2005: "Postscript to 'Deflationism, Meaning and Truth-Conditions'', in Beall,J.C. andArmour-Garb, B. eds 2005.

Bays, T. 2006: "Beth's Theorem and Deflationism", draft.

Beall, J.C. and Armour-Garb, B. 2001: "Can Deflationist be Dialetheist?" *Journal of Philosophical Logic* 30: 593–608.

Beall, J.C. 2001: "A Neglected Deflationist Approach to the Liar", *Analysis*, 61.2, pp. 126-9.

Beall, J.C. And Armour-Garb, B. eds 2005: *Deflationary Truth*, Chicago and La Salle, Open Court Press.

Beall, J.C. and Armour-Garb, B. eds 2006: *Deflationism and Paradox,* Oxford, Clarendon.

Bell, J. L.: 2016, "Infinitary Logic", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2016/entries/logic-infinitary/>

Blackburn S. and Simmons K. eds. 1999: *Truth,* Oxford

University Press.
Boghossian, P.A. 1990: "The Status of Content", *The Philosophical Review*, Vol. XCIX, No. 2.
Brandom, R.B. 1994: *Making it explicit: Reasoning, representing, and discursive commitment*. Cambridge,
Mass., Harvard University Press.
Blanshard, B. 1939: *The Nature of Thought*, George Allen and Unwin, London.
Boolos, G.S., Burgess, P.J., and Jeffrey, R.C. 2007: *Computability and Logic*, Cambridge, Mass., Cambridge
University Press.
Burgess, J. 1986: "The Truth is Never Simple", *Journal of Symbolic Logic* 51, 663–681.
Burgess, J. and Rosen, G. 1997: *A Subject with No Object: Strategies for Nominalistic Interpretation of
Mathematics*, Clarendon, Oxford.
Cantini, A. 1990: "A Theory of Truth Formally Equivalent to ID1" , *Journal of Symbolic Logic* 55, 244–59.
Cieslinski C. 2007: "Deflationism, Conservativeness and Maximality" *Journal of Philosophical Logic* 36: 695
705.
Cieslinski, C. 2010: "Truth, Conservativeness, and Provability", .*Mind*, 119, 409–22.
Cieslinski, C. 2015: 'The innocence of truth', *Dialectica*, 69(1), 61–85.
Cieslinski, C. 2017: *The Epistemic Lightness of Truth: Deflationism and its Logic*. Cambridge University Press.
Cieslinski, C.; Wcisło, B. & Łełyk, M. 2017: 'Models of PT- with Internal Induction for Total Formulae' *Review of Symbolic Logic* 10 (1):187-202.
Craig, W. and Robert V. 1958 , 'Finite axiomatizability using additional predicates', *Journal of Symbolic Logic* , 23 , 289 –308 .

Damnjanovic, N. and Stoljar, D. 2014. "The Deflationary Theory of Truth", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/fall2014/entries/truth-deflationary/>

David, M. 1994: *Correspondence and Disquotation: an Essay on the Nature of Truth*, Oxford University Press.

Davidson, D. 1990: "The Structure and Content of Truth", *Journal of Philosophy*, 87, pp. 279-328

Davidson, D. 1996: "The Folly of Trying to Define Truth". *Journal of Philosophy*, 93, pp. 263–78.

Davidson, D. 1984, *Inquiries into Truth and Interpretation*, Oxford University Press, Oxford.

Dummett, M 1959: "Truth", *Proceedings of the Aristotelian Society*, n.s. 59. Reprinted in Dummett, M. 1978:
Truth and Other Enigmas, Oxford, Clarendon Press.

Dummett, M.: 1963, "The Philosophical Significance of Gödel's Theorem", *Ratio* 5, 140– 155. Reprinted in Dummett 1978.

Dummett, M. 1978: *Truth and Other Enigmas*, Harvard University Press, Cambridge.

Dummett, M. 1990: "The Source of the Concept of Truth", in G. Boolos (ed.), *Meaning and Method: Essays in*
Honor of Hilary Putnam, Cambridge University Press, Cambridge. Reprinted in (Dummett
1993).

Dummett, M. 1993: *The Seas of Language*, Oxford University Press, Oxford.

Engström F. 2002: *Satisfaction Classes in Nonstandard Models of First-order Arithmetic*, Thesis for the Degree
of Licentiate of Philosophy, Department of Mathematics Chalmers University of Technology and Goteborg University.

Etchemendy, J.: 1988, 'Tarski on Truth and Logical Consequence', *Journal of Symbolic Logic* 53, 51–79.

Edwards, D. 2013; "Truth as a Substantive Property", *Australasian Journal of Philosophy* 91 (2):279-294.

Feferman, S. 1962: "Transfinite recursive progressions of axiomatic theories", *Journal of Symbolic Logic*, 27, pp. 259-316.

Feferman, S. 1964: "Systems of Predicative Analysis", *Journal of Symbolic Logic* 29, 1– 30.

Feferman, S. 1988: "Hilbert's Program Relativized: Proof-Theoretical and Foundational Reductions", *Journal of Symbolic Logic* 53, 364–384.

Feferman, S. 1991: "Reflecting on Incompleteness", *Journal of Symbolic Logic* 56, 1–49.

Feferman, S. 1992: "What Rests on What?", invited lecture, 15th International Wittgenstein Symposium: Philosophy of Mathematics, held in Kirchberg/Wechsel, Austria, 16-23 August 1992.

Field, H. 1980: *Science Without Numbers*, Princeton, Princeton University Press.

Field, H 1986: "The Deflationary Conception of Truth", in MacDonald, G. and Wright, C. eds *Fact, Science and Morality*, Oxford, Blackwell.

Field, H. 1992: "Critical Notice: Paul Horwich's 'Truth' '', *Philosophy of Science*, 59, pp. 321-30.

Field, H. 1994a: "Deflationist Views of Meaning and Content", *Mind*, Vol. 103, No.411, pp. 249-84.

Field, H. 1994b: "Disquotational Truth and Factually Defective Discourse", *Philosophical Review*, 103(3), pp. 405-52.

Field, H. 1999: "Deflating the Conservativeness Argument", *Journal of Philosophy*, 96, pp. 533-40.

Field, H. 2005: "Postscript to Deflationist Views of Meaning and Content", in Beall and Armour-Garb 2005.

Frege, G. 1918: "Thoughts", in his *Logical Investigations*, Oxford: Blackwell, 1977.

Friedman, H. and Sheard M., 1987: "An Axiomatic Approach to Self-Referential Truth", *Annals of Pure and*
Applied Logic 33, 1–21.
Gauker, C. 2001: "T-schema Deflationism versus Godel's First Incompleteness Theorem", *Analysis* 61, 129-135.
Glanzberg, M. 2001: "The Liar in Context", *Philosophical Studies* 103**,** 217–251.
Glanzberg, M. 2003: "Truth, Disquotation, and Expression" manuscript.
Glanzberg, M. (forthcoming): "A Contextual-Hierarchical Approach to Truth and the Liar Paradox", *Journal of*
Philosophical Logic.
Grzegorczyk, A. 2005, 'Undecidability without arithmetization', *Studia Logica*, 79 , 305 –313 .
Grover, D. 1992: *A prosentential theory of truth*, Princeton, NJ: Princeton University Press.
Grover, Camp, D.J. and Belnap N. 1975: "A Prosentential Theory of Truth", *Philosophical Studies* 27, 73–125.
Gupta, A. and Belnap, N. 1993: *The Revision Theory of Truth,* Cambridge, MA, MIT Press.
Gupta, A. 1993: "A Critique of Deflationism", *Philosophical Topics* 21, 57–81.
Gupta, A. 1982: "Truth and Paradox", *Journal of Philosophical Logic* 11, 1–60.
Gupta, A. 2006: "Do the Paradoxes Pose a Special Problem for Deflationists?", in Beall, J.C. and Armour-Garb,B. eds
Hájek, P. and Pavel P. 1993: *Metamathematics of First-Order Arithmetic*, Perspectives in Mathematical Logic, Springer,Berlin.
Halbach, V. 1994: "A System of Complete and Consistent Truth", *Notre Dame Journal of Formal Logic* 35, 311–327.
Halbach, V. 1995: "Tarski Hierarchies", *Erkenntnis* 43:

339-367.
Halbach, V. 1996: *Axiomatische Wahrheitstheorien*, Akademie Verlag, Berlin.
Halbach, V. 1999a: "Conservative Theories of Classical Truth", *Studia Logica* 62, 353–370.
Halbach, V. 1999b: "Disquotationalism and Infinite Conjunctions", *Mind* 108, 1–22.
Halbach, V. 1999c: "Disquotationalism Fortified", in A. Chapuis and A. Gupta (eds), *Circularity, Definitions, and Truth*, *Indian Council of Philosophical Research*, New Delhi.
Halbach V. 2000: "Truth and Reduction", *Erkenntnis* 53: 97–126.
Halbach, V. 2001a: "How Innocent is Deflationism?" *Synthese* 126: 167–194.
Halbach, V. 2001b : "Disquotational Truth and Analyticity" The Journal of Symbolic Logic, Vol. 66, No. 4, pp.19591973.
Halbach, V. 2002: "Modalized Disquotationalism", in proceedings of the conference, Principles of Truth. Truth,Necessity and Provability (Halbach, V. and Horsten, L. eds 2002).
Halbach, V. 2006: "How Not To State the T-sentences"Analysis 66 (2006), 276-280.
Halbach, V. (2009): "Reducing Compositional to Disquotational Truth",to appear on *The Journal of Symbolic Logic*.
Halbach, V. and Horsten, L., eds 2002: *Principles of Truth. Truth, Necessity and Provability,* proceedings of the conference.
Halbach, V. 2011: *Axiomatic Theories of Truth*. Cambridge: Cambridge University Press.
Heck, R. 2004: "Truth and Disquotation", *Synthese* 142: 317–352.
Heck, R. 2009:'The strength of truth-theories', draft.

Hempel, C., 1935: "On the Logical Positivists Theory of Truth", *Analysis* 2, 49-59.

Herzberger, H.: 1982: "Notes on Naive Semantics", *Journal of Philosophical Logic* 11, 61–102.

Hill, C. 2002: *Thought and World: An Austere Portrayal of Truth, Reference, and Semantic Correspondence.*

Cambridge, Cambridge University Press.

Hodges, W. 1997: *A Shorter Model Theory.* Cambridge: Cambridge University Press.

Horwich, P. 1998a: *Meaning*. Oxford, Clarendon.

Horwich, P. 1998b: *Truth.* Oxford, Blackwell (first edition 1990).

Horsten, L. 1995: "The Semantical Paradoxes, the Neutrality of Truth and the Neutrality of the Minimalist

Theory of Truth", in P. Cortois (ed.), *The Many Problems of Realism*, Vol. 3 of *Studies in the General Philosophy of Science*, Tilburg University Press, Tilburg, pp. 173–87.

Hyttinen, T. and Sandu, G. 2004: 'Deflationism and arithmetical truth', *Dialectica* 58 (3):413–426.

Jackson, F. , Oppy, G. and Smith, M. "Minimalism and Truth Aptness", *Mind*, Vol 103, No 411, pp. 287-302.

Joachim H.H., 1906: *The Nature of Truth, Oxford* University Press, Cambridge.

Kaufmann, M. 1977:"A Rather Classless Model" *Proceedings of the American Mathematical Society*, Vol. 62,

No. 2 pp. 330-333

Kaye, R. 1991: *Models of Peano Arithmetic*, Oxford Logic Guides, Oxford University Press.

Ketland, J., 1999: "Deflationism and Tarski's paradise", *Mind* 108.

Ketland, J.2000: "Conservativeness and translation-dependent T-schemes", *Analysis* 60, pp. 319–28.

Ketland, J. 2000b: "A proof of the (strengthened)

Liar formula in a semantical extension of Peano arithmetic".
Analysis 60: 1–4.
Ketland, J. 2005: "Deflationism and the Gödel phenomena: reply to Tennant", *Mind* 114, 75Y88.
Kirkham, R.L. 1992: *Theories of Truth*, Cambridge, MA, MIT Press.
Kossak R 1985: "A Note on Satisfaction classes", *Notre Dame Journal of Formhal Logic* Volume 26, Number 1.
Kotlarski H. 1991: "Full Satisfaction Classes: A Survey", *Notre Dame Journal of Formal Logic* Volume 32,
Number 4.
Kotlarski, H., Krajewski, S. and Lachlan, A. 1981: "Construction of Satisfaction Classes for Nonstandard
Models", *Canadian Mathematical Bulletin* 24, 283–93.
Kripke, S. 1975: "Outline of a Theory of Truth", *Journal of Philosophy* 72, 690–712.
Künne, W. 2003: *Conceptions of Truth,* Oxford, Clarendon.
Lachlan, A.: 1981: "Full Satisfaction Classes and Recursive Saturation", *Canadian Mathematical Bulletin* 24, 295–297.
Leeds, S. 1978: "Theories of Truth and Reference", *Erkenntnis*, 13, 111-129.
Maudlin, T. 2004: *Truth and Paradox: Solving the Riddles.* Oxford, Clarendon.
Leigh, G. E. & Nicolai, C. 2013: 'Axiomatic truth, syntax and metatheoretic reasoning', *Review of Symbolic Logic* 6 (4):613-636.
Łełyk, M. and Wcisło, B. 2017: 'Models of weak theories of truth'. *Archive for Mathematical Logic* 56 (5-6):453-474.
Łełyk, M. and Wcisło, B. 2019: 'Models of positive truth'. *Review of Symbolic Logic* 12 (1):144-172.
McGee, V. 1985: "How Truthlike can a Predicate Be?", *Journal of Philosophical Logic* 14, 399–410.

McGee, V. 1991: *Truth, Vagueness and Paradox*, Hackett, Indianapolis, IN.

McGee, V: 1992: "Maximal Consistent Sets of Instances of Tarski's Schema (T)*", *Journal of Philosophical Logic* 21: 235-241.

McGrath, M. 2000: *Between Deflationism and Correspondence*. New York, Garland Publishing.

Murzi, J. & Rossi, L. 2020. Conservative deflationism? *Philosophical Studies* 177 (2):535-549.

Nicolai, C. & Piazza, M. 2019. The Implicit Commitment of Arithmetical Theories and Its Semantic Core. *Erkenntnis* 84 (4):913-937.

Neurath, O. 1983: *Philosophical Papers 1913-46*, eds. Robert S. Cohen and Marie Neurath, D. Reidel Dordrecht and Boston.

Parsons, C. 1983: *Mathematics in Philosophy*, Cornell University Press, Ithaca, Chapter

Peirce, C.S. *Collected Papers*:, 8 vols. Edited by Charles hartshorne, Paul Weiss and Arthur Burks, Harvard University Press, Cambridge, Massachusetts, 1931-1958.

Picollo, L. and Schindler, T. 2018. Disquotation and Infinite Conjunctions. *Erkenntnis* 83 (5):899-928.

Priest, G. 1987: *In Contradiction* (Martinus Nijhoff, Dordrecht).

Putnam, H. 1971:*Philosophy of Logic*, New york, reprinted in H. Putnam 1979 *Mathematics Matter and Method. Philosophical Papers.* Vol. I. Second edition, Cambridge, Cambridge University Press.

Putnam, H. 1978: *Meaning and the Moral Sciences*, Routledge, 1978.

Putnam, H. 1981: *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W.V. 1970: *Philosophy of Logic.* Englewood Cliffs, Prentice Hall.

Quine W.V. 1948: "On What There is", reprinted in Quine 1980.

Quine W.V. 1980: *From a Logical Point of View: Nine Logico-Philosophical essays*. Second edition revised Cambridge MA; Harvard University Press.

Quine W.V. 1995: *From Stimulus to Science*, Cambridge, MA, Harvard University Press.

Raatikainen P. 2002: "Deflationism and Gödel's theorem – a comment on Gauker" *Analysis* 62.1, January 2002, pp. 85-87.

Raatikainen *P.* 2003: "Hilbert's Program Revisited" *Synthese* 137: 157–177, .

Raatikainen P. 2005: "On Horwich's way out" *Analysis* 62.1, January 2002, pp. 85–87.

Ramsey, F.P. 1927: "Facts and Propositions", *Proceedings of the Aristotelian Society*, Vol.7.

Reinhardt, W.1986: "Some Remarks on Extending and Interpreting Theories with a Partial Predicate for Truth",

Journal of Philosophical Logic 15: 219-251.

Rescher, N., 1973: *The Coherence Theory of Truth*, Oxford University Press, Oxford

Resnik, M. 1990: "Immanent truth", *Mind* 99, 405–424.

Restall, G. 2006: 'Minimalists Can (and Should) Be Epistemicists, and it Helps if They Are Revision Theorists

Too', in Beall, J.C. and Armour-Garb, B. eds 2006.

Robinson. A, 1963: "On languages which are based on non-standard arithmetic", *Nagoya Math. J.*, 22:83–117.

Schindler, T. 2015: 'A Disquotational Theory of Truth as Strong as Z 2 –' *Journal of Philosophical Logic* 44 (4):395-410.

Shapiro,S.1991: *Fondation without Foundationalism: a Case for Second Order Logic,* Oxford University Press, Oxford.

Shapiro, S. 1998: "Proof and Truth: Through Thick and Thin", *Journal of Philosophy*, 95, pp. 493-521

Shapiro, S. 2002: "Deflation and Conservation" in proceedings of the conference, *Principles of Truth. Truth,*

Necessity and Provability (Halbach, V. and Horsten, L. eds 2002).

Sheard M.1994: "A Guide to the Truth Predicate in the Modern Era" *Journal of Symbolic Logic* Vol. 59, No. 3.

Sheard M. 2001: Weak and Strong Theories of Truth" Studia Logica 68: 89–101.

Schiffer, S. 2003: *The Things We Mean*, Clarendon Press, Oxford

Schütte, K. 1977: *Proof Theory*, Springer, Berlin.

Simmons K. 1999 "Deflationary Truth and the Liar" *Journal of Philosophical Logic* 28: 455–488.

Simpson, S. 1998: Subsystems of Second Order Arithmetic, Springer, Berlin.

Strollo, A. 2014a: "Deflationism and the Invisible Power of Truth", *Dialectica* Vol. 67, N° 4 , pp. 521–543.

Strollo, A. 2014b: "How simple is the simplicity of truth? Reconciling the Metaphysics and the Mathematics of Truth", in *New Frontiers in Truth*, Cambridge Scholars Publishing. pp.161-175.

Strollo, A .2018: "Making sense of Deflationism: conservativity and interpretability", in *Truth, Existence, and Explanation*, (Springer), eds. M.Piazza, G. Pulcini,. pp. 89-105.

Soames, S. 1997: 'The Truth about Deflationism', in Villanueva (ed.), *Truth*, Vol. 8 of *Philosophical Issues*,

Ridgeview, Atascadero, pp. 1–44.

Soames, S. 1999: *Understanding Truth.* Oxford, Oxford University Press.

Strawson, P. 1949: "Truth", *Analysis* 9, 83–97.

Takeuti, G.: 1987, *Proof Theory*, 2nd edn, North Holland,

Amsterdam.
Tarski, A. 1956: "The Concept of Truth in Formalized Languages", *Logic, Semantics, Metamathematics*,
Clarendon Press, Oxford, pp. 152–278.
Tennant, N., 2002: "Deflationism and Gödel Phenomena", *Mind 111*, pp. 551-582.
Tennant, N., 2004: "Deflationism and Gödel Phenomena: reply to Ketland", *Mind 114*, pp. 89-96.
Tennant, N. 2010: "Deflationism and the Godel Phenomena: Reply to Cieslinski", *Mind* 119 (474):437-450.
Visser, A. 1989: "Semantics and the Liar Paradox", in *Handbook of Philosophical Logic*, Vol. 4. D. Gabbay and
F. Guenthner, eds., (D.Reidel, Dordrecht), 617-706.
Waxman, D. 2017. Deflationism, Arithmetic, and the Argument from Conservativeness. *Mind* 126 (502):429-463.
Weir A: "Ultramaximalist Minimalism!" *Analysis*, Vol. 56, No. 1, (Jan., 1996), pp. 10-22
Williams, C.J.F., 1976: *What is Truth?*, Cambridge University Press, Cambridge.
Williams, M., 1986: "Do We (Epistemologists) Need a Theory of Truth?", *Philosophical Topics* XIV
Wittgenstein L. 1956: "Remarks on the Foundations of Mathematics", ed. G.H. Von Wright, R. Hees, G.E. M.
Anscombe, trans. G.E.M. Anscombe, Blackwell.
Wright, C. 1992: *Truth and Objectivity*, Cambridge, Mass., Harvard University Press.
Yablo, S. 1985: 'Truth and Reflection', *Journal of Philosophical Logic*, 14, pp. 297-349.
Yablo, S. 1993: 'Paradox Without Self-Reference', *Analysis*, 53, pp. 251-2.

While the nature of truth is a classical philosophical conundrum, deflationism holds that truth is an insubstantial property and its predicate only serves expressive purposes. Accordingly, it promises to solve the mystery of truth by dissolving it in its logic. But is deflationism an adequate theory of truth? This book contributes to assess the deflationary conception by discussing it in the context of axiomatic theories of truth. In particular, it offers a comprehensive and critical discussion of deflationism from the point of view of the so called *conservativeness argument* and its interplay with formal theories. The book also provides a neat example of the fruitful application of logic-mathematical methods to shed light on philosophical issues.

Andrea Strollo is Associate Professor at the Department of Philosophy of Nanjing University (China). His main areas of research are in logic and philosophy of logic, with a specific focus on the notion of truth.

ISBN 978-88-6938-241-3

€ 16,00